FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

On Estimating Congestion Causes from Floating Car Data using White-Box Classifiers

Márcio Rui Pereira Coelho

Master in Information Engineering

Supervisor: Professor Pedro M. d'Orey

Co-Supervisor: Professor Ana Aguiar

October 18, 2017

On Estimating Congestion Causes from Floating Car Data using White-Box Classifiers

Márcio Rui Pereira Coelho

Master in Information Engineering

October 18, 2017

Abstract

Traffic congestion estimation in a city is an important problem which is useful to public transportation networks and even particular drivers. Its effects have been evaluated on 0.8% of Europe's GDP and that value is only going to increase. This thesis proposes a solution for said problem by using white-box classifiers. These allow the retrieval and interpretation of more information when compared to the more standard approach of stronger prediction methods. Starting on data collected by floating cars (taxis) a series of filters and discretizations are applied, re-sampling and aggregating the information. Using Classification and Regression Trees and evaluating the results in both prediction and interpretation using specific metrics, it is shown that information can be retrieved from the decision tress, by doing a statistical analysis on the feature importance as well as the thresholds of each feature that appear on said trees. The results display the hours at which congestion occurs in the center of the city and that other cells hold more importance when predicting the traffic state of a specific cell. Also, we show that the peripheral zones of the urban center provide more information to the traffic congestion estimation, as well as the main roads that lead to the urban center are more important than the roads in the urban center.

Acknowledgements

I would like to express my sincere appreciation to my advisors, Pedro M. D'Orey and Ana Aguiar, who offered invaluable support and helpful comments. Without their guidance this project wouldn't be possible at all and I'm really thankful for giving me the chance to pursue this project. I'd also like to thank Luís Moreira-Matias for his input and opinion on the different stages of this work.

Secondly, a very special thanks must go to my former students and colleagues. They helped me and put up with the days where my patience wasn't where it should, without a single complaint.

Further, to my friends and family, for always staying by my side, even when i barely had time for them.

Last but not least, my wife. She is the main reason why i started this master's degree and has always been there for me, specially in my most stressful times. For believing in me from day one, having my back, and for letting me know when i was over thinking things.

To all, a most sincere Thank You.

Márcio Coelho

"Winners are not afraid of losing. But losers are. Failure is part of the process of success. People who avoid failure also avoid success."

Robert T. Kiyosaki, Rich Dad, Poor Dad

Contents

Abbreviations

FCD	Floating Car Data
GPS	Global Positioning System
MFD	Macroscopic Fundamental Diagram
LOS	Level of Service
CDF	Cumulative Distribution Function
CART	Classification and Regression Trees
RF	Random Forest
IQR	Inter-Quartile Range
SCV	Stratified Cross-Validation
CV	Cross-Validation
TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives
mF1	Micro F1
aF1	Average F1
Mf1	Macro F1

Chapter 1

Introduction

1.1 Motivation

Urban traffic congestion is a major problem that affects not only the people's daily routines, but also a city's economic development. The effect of congestion has been evaluated in 19 different European cities to an estimated 183 €billion cost throughout the next decade (0.8% of Europe's GDP), reflecting on the continent's sustainable growth, and is expected to grow up to 300 €billion by 2030 (Debyser, 2014). Drivers in Belgium or Netherlands waste on average over 50 hours per year in traffic congestion (Figure 1.1) and that value is only going to increase if nothing is changed.

Figure 1.1: Average Number of Hours 2012 vs. 2013.

It is then urgent to find a way to reduce the traffic congestion. But this proves harder than expected, as urban traffic flow is complex and constantly changing, making it difficult to acquire the current and to predict the future traffic conditions.

There are at least two major problems that need to be solved in order to estimate traffic congestion as well as discover its causes. First, from where do you retrieve the information required and how do you process it? Floating Cars (or mobile probes), or more specifically, taxis equipped with Global Position System (GPS), are becoming an increasingly common way to collect real-time traffic flow in a large-scale road network when compared to other methods (Zheng, 2015). But processing has its own problems like low GPS accuracy, different mobility patterns for each city, or means to deal with the different taxi operations (Castro et al., 2013b). Also, there are

methods in which data from different sources is used which is a challenge by itself (Ambühl and Menendez, 2016). Secondly, how do you infer which causes are more important or have a higher weight in increasing traffic congestion? While this project could focus on predicting congestion by using algorithms stronger suited for higher interpretability, the focus will instead be on higher interpretability. Thus, the trade-off is between the accuracy in predicting congestion versus the amount of information one can obtain. This project will focus on the latter one by using white-box classifiers, which allows the inference of features' importance, giving a larger understanding on the congestion mechanics, its propagation and starting causes.

1.2 Challenges

Even though public transportation fleets are increasing the number of GPS equipped vehicles, there are inherent challenges to using this a means to retrieve information. They are:

- Data Sparsity - on the specific case of taxis being used as mobile probes, there are cases where more information is collected near the taxi stands or near tourist points, which might reduce the coverage of the taxis throughout the network. On a more generic case, this causes information to be focused on certain roads or even periods of day;
- Data Quality - on an attempt to sometimes cut costs, companies might resort to low-cost sensors that typically have lower quality, directly influencing the quality of the collected data;
- Missing Data - occasionally, GPS data might not be properly transmitted/received, or the taximeter in the taxi case might fail, causing a lapse in the data collected that can span minutes or sometimes hours. This can be caused by GPS device's errors, lack of signal due to location, etc.;
- External Factors - bus lanes, construction sites or accidents, represent factors which are not directly related to the collected data but might influence the estimation of traffic;

This project will attempt to minimize the impacts of each one of these problems by preprocessing the data in order to remove erroneous information and focusing on the main areas of the city where more information is available. Also, small gaps of missing data are corrected with the imputation of data. As for external factors, it has been left for future work.

1.3 Contribution

The focus and contribution of this project will be to understand if traffic congestion can be predicted from interpretability instead of focusing on prediction. The goal is to understand how congestion propagates and its source in urban environments, as well as to understand how externalities (e.g. weather) affect the traffic flow in urban environments, by using white-box classifiers like such as decision trees. This work argues that by using white-box classifiers, it is possible

to retrieve such information, using data from a taxi's fleet, preprocessed to remove outliers and filtered. Three types of discretization are then applied: space discretization which creates a grid of squared cells on the city of Porto, allowing the independent study of each cell; time discretization which samples the data in time slots of 15 minutes to reduce the amount of data at hand, allowing for better processing times; state discretization to classify each velocity in one three possible traffic movement states. Following these steps, a dataset is generated with added information such as weather conditions and speed and acceleration statistics. These will be some of the features used in the different created decision tree models. In the end, each model will be evaluated both in terms of prediction and in terms of interpretability, by studying the generated decision trees as well as the importance of each selected feature.

1.4 Thesis Structure

Chapter 2: Literature Review - This chapter delivers an overview of the current techniques for the estimation of traffic. First, a group of techniques and methods for the collection of traffic information is presented and then current methods for estimating traffic information is shown and detailed.

Chapter 3: Exploratory Data Analysis - This chapter presents an analysis to the data at hand. It starts with a description of the scenario of the city and it's population, habits, logistic, traffic and weather. Also described is the Floating Car Data technique as well as the raw data collected during the month of April, 2016. After the weather data description, the data is explored and described in terms of number of active taxis and trips, as well as velocity and its distribution and behavior.

Chapter 4: Methodology - This chapter describes the steps taken to reach the proposed objectives. The chapter starts by explaining the preprocessing required to "clean" the data. Next, discretizations in space, time and state are analyzed, as well as their reasoning. Lastly, different classification models such as hierarchical or weekday based, are used and developed on this project are explained. In the end, an explanation of the metrics used to classify each model is given.

Chapter 5: Results and Discussion - In this chapter the results for each model are displayed, both in terms of predicting capabilities and interpretation. Results for the different metrics are shown as well as samples for the different features per model and respective decision trees.

Chapter 2

Literature Review

This chapter delivers an overview of the current techniques for the estimation of traffic. First, a group of techniques and methods for the collection of traffic information is presented and then current methods for estimating traffic information is shown and detailed.

2.1 Traffic data collection methods

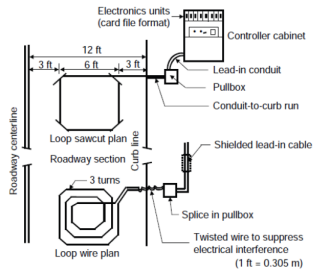
Traffic detection is nowadays an important component of the intelligent transportation systems, since it can provide information on historical and real-time traffic data. Such data can be used for traffic control and management or transportation planning and optimization, among others.

Different traffic estimation techniques were developed throughout the years, answering the increasing needs for diverse data, but no single approach has provided all the necessary data. In this chapter the main data collection techniques are reviewed both in terms of mechanisms and applications. There are currently three types of traffic detection technologies (Cheung and Varaiya, 2006) based on the sensor's location:

- *Intrusive* - any sensor that is built within or across the pavement (e.g.: Inductive Loop Detection and Weight-In-Motion);
- *Non-Intrusive* - any sensor built above or by the road (e.g.: Infrared-Based and Video-Image Processing);
- *Off-Roadway* - these sensors do not need any type of equipment built on the road or by it (e.g.: Automatic Vehicle Identification and Probe Vehicle with Global Positioning System (GPS));

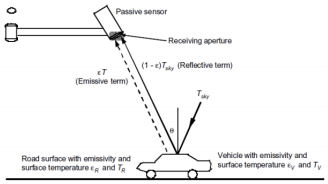
2.1.1 Intrusive Detection Technologies

These methods require the installation of the sensor directly onto or into the road surface and are usually installed directly on the pavement surface in either saw-cuts in the road surface, by tunneling under the surface or by anchoring directly to the pavement surface. Measured metrics can be



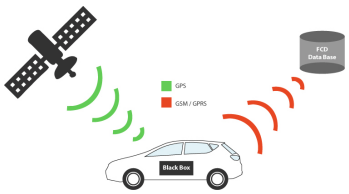
(a) Scheme for ILD's operation - Source: Klein et al. (2006)

(b) Weight in Motion system seen from the outside - Source: *Aditya Technologies* [1]

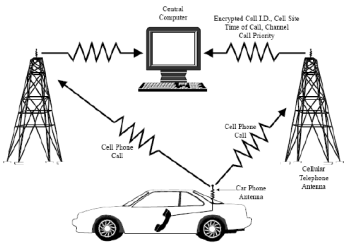


(c) Emission and reflection of energy by a vehicle - Source: Mimbela and Klein (2000)

(d) Exemplification of Video-Image Processing Camera - Source: *KEYTOP* [2]



(e) FCD Illustration - Source: *TPS* [3]



(f) Probe Vehicle with Mobile Phones exemplification - Source: Leduc (2008)

Figure 2.1: Examples of Traffic Data Collection.

the count of passing vehicles, or just the presence of one. The main advantage lays on the great accuracy these methods can achieve when detecting traffic. Drawbacks include the disruption of traffic for installation and repair of the equipment, which can increase their overall cost (Mimbela and Klein, 2000).

Inductive Loop Detector (ILD)

It is one of the most used traffic detection technology (Klein et al., 2006) and it requires an installation under the pavement. The system is shown in Figure 2.1 (a). It is possible to see the wire loop which is under the road. This loop is electrically powered and causes a magnetic field. Whenever a vehicles passes or stops by the loop, its inductance is reduced, which sends a pulse to the traffic signal controller. If the pulse's frequency is over a pre-specified threshold, then a vehicle passage is detected (Cheung and Varaiya, 2006).

Weight-In-Motion (WIM)

Weight-In-Motion systems are capable of estimating the gross weight of a vehicle even at speeds ranging from $16km/h$ up to $130km/h$ (ASTM, 2002). It is widely used for weight limitation and tolling, because it does not require vehicles to pull over and test their weights. It is also useful for bridge design and monitoring (Moses, 1979).

The WIM system is usually composed of two main parts: one which provides the power for the system to collect and store data, while the other consists of sensors and cables to transmit data. On the outside, the appearance is very similar to an ILD (Figure 2.1 (b))

2.1.2 Non-Intrusive Detection Technologies

Non-intrusive methods refer to methods which do not require any installation directly on the road, but instead above or by it. The advantages of these methods are the easy maintenance and installation with little to no traffic disruption in some cases. Accuracy can also be quite high when compared to intrusive methods (Minge et al., 2010) and it can also provide other metrics such as vehicle speed, type classification and lane distinction. The disadvantages of such methods is the sometimes hard to configure and processing of data (Mimbela and Klein, 2000).

Infrared-Based System (IBS)

Infrared-Based systems are usually built by the road side in order to detect approaching traffic. They can be used to measure vehicular volume or speed. They are divided in two groups: Passive and Active (Klein et al., 2006). Active Sensors: these sensors are constantly emitting radiation and measuring the time it takes to receive the reflection of it. If the difference between received and sent time is shorter than the typical value, then a vehicle has passed. Passive Sensors: sensors that detect the energy that is emitted from vehicles, road surfaces or any object inside the detection range. This energy is then converted into electrical signals and processed and analyzed to detect the present of a vehicle. An example of the system can be found in Figure 2.1 (c) where it is possible to see the emission of energy sent from a vehicle as well as its reflection, both detected by the sensor.

Video-image Processing (VIP)

Video-image detectors are used for roadway surveillance due to their ability to transmit closed circuit television imagery to a human operator. Nowadays, though, that analysis is done automatically. The system analyzes the traffic detection zone and detects changes between successive frames by analyzing the variation of gray levels in the video frames. They are usually composed by three parts: cameras for video recording (Figure 2.1 (d)), processor for image processing and software for imagery analysis and traffic detection (Mimbela and Klein, 2000).

2.1.3 Off-roadway Detection Technologies

Off-roadway detection technologies refer to those that do not need any hardware to be built on the road or next to it. The metrics that can be typically measured by this type of detectors are vehicle location, lane distinction, traffic direction, speed, etc. The main advantages are lack of on the road installation of any sorts and easiness to install on the chosen vehicle. Also, since it is a mobile

detector, it covers areas beyond the installation place, unlike the previous methods. The disadvantages are the sometimes loss of GPS network as well as location errors which can invalidate collected data (Cheung and Varaiya, 2006). Also, data protection concerns can sometimes cause privacy concerns.

Probe Vehicle with GPS - Floating Car Data (FCD)

Vehicles equipped with a GPS system, either via mobile phones or incorporated on the vehicle, are considered mobile probes. As they drive through the road network, their location and speed information can be calculated and stored (Figure 2.1 (e)), or in some cases, consulted in real-time (Schäfer et al., 2002). It is emerging as a reliable and cost-effective way to gather traffic data for a large road network and to improve predictions of traffic state (C.Fabritiis et al., 2008).

Using this system, it is possible to create applications for taxi and bus networks, which allow users to know the location of these means of transportation as well as their estimated time of arrival at the respective stops (Zhu et al., 2011). Also, taxi companies are increasing the number of vehicles that are equipped with a GPS receiver. They make use of the available communications infrastructures (e.g. UMTS/LTE) to transmit their position in real time, to the taxi dispatching system. It is a powerful technique to assess traffic conditions in urban areas given that a large number of vehicles collect such data, taxis for example, giving an *in loco* state of the traffic.

Probe Vehicle with Mobile Phones

This technique is similar to FCD but with some structural differences: instead of using satellites, phone antenna base stations are used, and GPS receivers are replaced by mobile phones (Figure 2.1 (f)). Also, there are two methods to find the location of a mobile phone: either by triangulation, where at least three antennas are used to find the location, or from mobile phones that already include a GPS. Due to the higher penetration rate of mobile phones, at least one can be found in a vehicle, unlike GPS systems. This compensates the lower accuracy ($100m$) when compared to the GPS counterpart ($10m$). This method can provide a great percentage of coverage if deployed nationally, but privacy concerns are raised by the public about unauthorized use of information making it not suitable for a large scale deployment.

2.2 Traffic Estimation

With the use of the different techniques mentioned before, all sorts of information (car speed, traffic count, traffic flow, density, to name a few) can be retrieved and processed which can lead to new and improved information like traffic states (Seo et al., 2017), delays in public transportation (Abdelfattah and Khan, 1998), decrease in pollution (Stevanovic et al., 2009) or optimization of traffic in a city (Mirchandani and Head, 2001). These advantages have the potential to improve the overall life quality of every user.

Among some of most traditional techniques is the usage of ILD (Kwon et al. (2003), Jeng and Chu (2014)) and VIP (Li et al., 2013). These techniques can easily provide real-time traffic measures like the number of vehicles moving in covered roads, but due to high costs of both installation and maintenance, the usage of these methods is restricted to a city's main road, which typically represents only a fraction of the full network. Certain models can be used to infer the behavior for the rest of the network (Kwon and Murphy, 2000) but typically require too many parameters, which imply an increase in both complexity and time consumption to collect enough data to estimate all of them if the model were to retain a high performance, hence why these models usually have a decrease in performance.

Another common method is the use of Fundamental Diagrams (FD) ((Derrmann et al., 2017), (Saberi et al., 2014), (Ambühl et al., 2017)). These represent the correlation from various simple metrics like speed, density or flow, which are used to infer other more complex measures like congestion, density or state. The largest downside of such method is the big data requirement per road in order to truly characterize it, not to mention that FD tend to have too much noise in urban areas, and are better suited for highway roads.

Novel methods include the usage of social media to predict traffic volume (Wang et al., 2015), but are often constrained by a heterogeneous distribution of users in the network, lacking representativeness of the overall traffic.

2.2.1 Traffic Estimation using FCD

Several previous works like C.Fabritiis et al. (2008), Kong et al. (2016) and Zhan et al. (2017) have dealt with the specific case of estimating traffic using FCD. They developed systems which take advantage of the higher coverage and great amount of information. Their approaches are listed in table 2.1 and described after.

	Traffic Estimation Metric:				Machine Learning Method			Techniques				
Article	Congestion	Volume	State	Speed	ANN	Fuzzy Logic	KNN	Fundamental Diagram	Map-Matching	Road Classes	Week/Weekdays Distinction	Performance
C.Fabritiis et al. (2008)				✓	✓				✓			2∼18% (MAE)
Zhan et al. (2017)		✓	✓	✓			✓	✓	✓	✓	✓	23% (MAE)
Kong et al. (2016)	✓	✓		✓		✓			✓			85% (accuracy)

Table 2.1: List of articles and their Goals, Methods and adopted techniques.

Metrics: Even though data was collected with the same technique, different metrics were estimated, like traffic congestion, volume, state or speed. Traffic congestion, estimated by Kong et al. (2016), refers to a condition on road networks occurring when the use of such network increases, causing a decrease in the velocity of vehicles in the road. Characteristics of such situation include slower speeds, longer trip times and increased waiting lines. Traffic volume, estimated by both Zhan et al. (2017) and Kong et al. (2016), is the amount of vehicles moving on the roads at a particular section, during a particular time. Traffic states, estimated by Zhan et al. (2017), indicate the level of service for each road, depending on the amount of cars that are passing by a certain section at a certain time, as well as their speed. It can be separated in three main classes: Free Flow (fastest state, traffic movement without any constraint), Synchronized Flow (middle state, where the traffic movement is slightly restricted) and Congestion (slowest state with traffic movement

reaching speeds near zero). Further details about each class will be described in section 4.3.3. Lastly, traffic speed, estimated by all the mentioned articles, is the distance covered per unit of time. Since only the floating car data is available, only the speed of that car or traffic volume of the network is known, while the speed of the other vehicles in the road can only be estimated.

Methods: The approaches for such goals can be quite different. Some of the most used techniques are Artificial Neural Networks (ANN), fuzzy methods, spatio-temporal pattern matching and K-nearest neighbor. ANN (used by C.Fabritiis et al. (2008)) are quantitative models that learn to associate input and output patterns adaptively with the use of learning algorithms without understanding the fundamental or physical relationships between them. They are composed by multiple nodes which imitate biological neurons of the human brain. Each node takes the input data and performs simple operations on the data and the result is passed to the following nodes.

Fuzzy comprehensive models, used in Kong et al. (2016), is based on fuzzy set theory developed by Zadeh (1965) and it focus on capturing the uncertainties inherent in a system. This is achieved by a series of membership functions which determine the relationship between factor sets (information such as traffic volume or speed) and evaluation sets (traffic states).

K-nearest neighbor, as seen in Zhan et al. (2017), is used as a means to infer road features such as speeds, by searching on the nearest neighbors (other roads with similar speed profiles) to the road that needs to be classified, and using the average of the neighbor's values.

Techniques: Different techniques can be used that allow the further retrieval of more information or different aggregation of data. Some approaches make use of Fundamental Diagrams, Map-Matching, Road Classes and Week/Weekdays Distinction. *Fundamental Diagrams*, used in Zhan et al. (2017) describe the empirical relationship between traffic volume, density and speed. These diagrams, first derived by Greenshields et al. (1935), can provide important parameters that characterize specific roads or groups of similar roads: capacity volume (maximum volume of vehicles on a road under optimum traffic conditions), critical speed (speed corresponding to capacity volume) and free flow speed (maximum velocity achieved in a road, while under the legal speed limit). They can also serve as means to classify traffic into the states mentioned before.

Another useful technique is *Map-Matching* (C.Fabritiis et al. (2008), Kong et al. (2016) and Zhan et al. (2017)), which allows coordinates collected by the floating cars to be matched to the road network representation. This is used to complement the collected information with, for example, road speed limits or the types of roads traveled.

Road classes and *week/weekdays* distinction are two ways to aggregate information, either by grouping roads with the same class (e.g.: highways or urban roads) or by organizing the data in two groups: week days or weekend days. Grouping road classes (Zhan et al. (2017)) can be beneficial to traffic estimation since different road classes have different speed profiles, retaining completely different characteristics. Separating between week days (Zhan et al. (2017)) and weekend days is also a common technique as these groups of days usually have very different traffic behaviors.

In terms of results, the available articles can reach 2% of mean absolute error (MAE) when predicting *speed* (C.Fabritiis et al., 2008), or 23% for (Zhan et al., 2017). When predicting congestion, the accuracy can go as high as 85% (Kong et al., 2016). These are very good results with

the aim on prediction, but retrieve little to no information regarding how the predicted metrics behave or are propagated throughout the studied area.

These articles aim to improve the prediction of either traffic states, speeds or volume, by using robust methods that are optimized in that direction. What this project proposes is to switch the aim from prediction to interpretability. This will be achieved by predicting the traffic states of a road network, but also what causes such traffic states, and what influences them the most. To achieve that, machine learning methods more oriented to the retrieval of information will be used, rather than focusing on methods that are more suited to prediction, like ANN or Fuzzy approaches. Also, a spatial discretization is implemented which should allow to focus on small cells instead of the full network.

Chapter 3

Exploratory Data Analysis

This chapter provides an analysis of the data at hand. Starting with a description of the scenario of the city and its population, habits, logistic, traffic and weather. Also described is the Floating Car Data technique as well as the all the data it collected during the month of April. After the weather data description, the data is explored and described in terms of number of active taxis and trips, as well as velocity and its distribution and behavior.

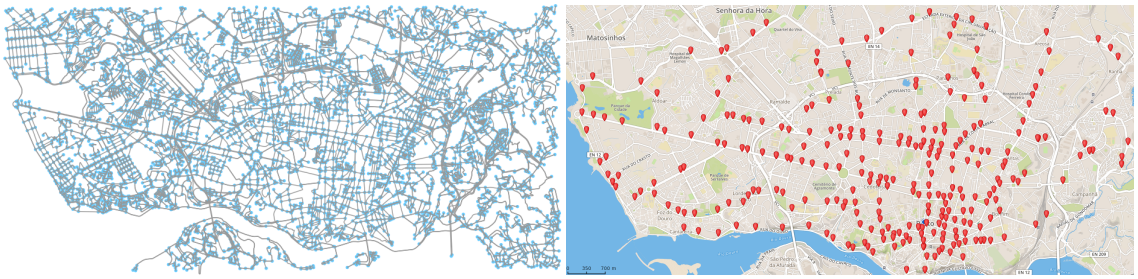
3.1 Scenario

The dataset made available for this thesis was collected in the city of Porto, which is the second largest city in Portugal. The metropolitan area of Porto extends beyond the administrative limits of the city and has a population of over 1.7 millions in an area of over 2000 km^2 [1]. The bulk of the traffic happens in the early morning when the population first enters the city through one of the many bridges available or come from the neighboring areas, and in the afternoon/evening when they leave it. The city itself has an area of $41.4km^2$ but it concentrates a lot of attractions which increase the number of tourists that visit the city each year [2] explaining why the city's traffic remains active in the summer season. Its road network has a length of 965 km and it is mainly composed by lower capacity ways with important limited access highways transversing the city. Around 16% of the city's intersections are signalized (Ferreira and d'Orey, 2012).

Looking at Figures 3.1 (a) and (b) it is possible to see the complex distribution of the road network, as well as the coverage of traffic lights throughout the city, which increase in the downtown area, continuing through main roads, but decreasing further away from downtown.

[1] http://portal.amp.pt/pt/

[2] http://observador.pt/2017/01/19/turismo-do-porto-fecha-2016-com-68-milhoes-de-dormidas-e-quase-atinge-meta-para-2020/



(a) (b)

Figure 3.1: Roads map of the city of Porto and its semaphore - Higher Concentration of semaphores in the downtown area and along main roads.

3.2 Data Acquisition

3.2.1 Floating Car Data

Floating car data was collected by the taxi dispatch system of a fleet of more than 440 vehicles in the city of Porto, Portugal, for the month of April, 2016, which allowed this project to be compared with other ongoing projects that focus on the same time period. Table 3.1 describes the basic features of the data set collected. The total number of 441 taxis generated almost 120 million points with a total of 283000 trips. The average duration of each trip is less than 8 minutes and the distance traveled is in average 6.6km. This shows that taxis are mainly used for small trips to specific locations (e.g. hospitals), usually to make the connection between other types of public transportations by tourists.

Metrics	Unit	Value
Evaluation Period		1-30 April 2016
Number of Taxis		441
Number of GPS points		119225379
Number of Trips		283720
Average Trip Length	m	6623
Average Travel Time	s	441
Average Travel Speed	km/h	18.5

Table 3.1: Taxi trajectories dataset properties.

FCD collects both the position and timing of the vehicle as well as taxi state information (busy, free, pickup, etc.), resorting to GPS technology and taximeter, respectively. Positioning information is acquired with a frequency of 1 Hz. The taxi state is acquired during state transitions (e.g. taximeter activation at the end of a service). A vehicle trajectory is composed of a sequence of ordered timestamped geospatial coordinates - latitude, longitude. Each trip is composed of:

- Trip Identifier - an 18 digit unique number to each trip
- Taxi Identifier - an 8 digit unique number for each taxi
- Taxi State - one of the following 6 states:

  - *Busy* (with an on board passenger)
  - *Free* (without an on board passenger after a service)
  - *Pickup* (en route to pick a new passenger, private call)
  - *Central Pickup* (en route to pick a new passenger, central call)
  - *Taxi Stand* (stopped at a taxi stand)
  - *Break* (driver on a break)

- Trip Start Time - a timestamp of the start of the trip using *UNIX time*
- Trip End Time - a timestamp of the end of the trip using *UNIX time*
- Vehicle Trajectory - a sequence of latitude and longitude positions

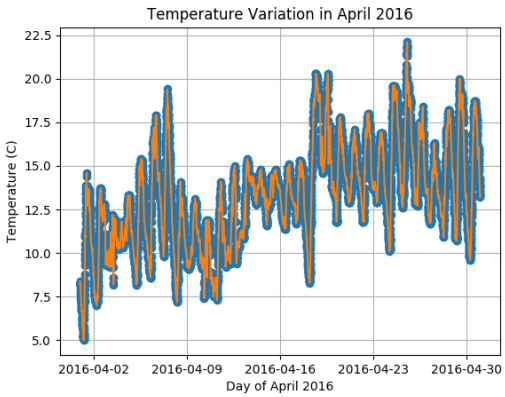
This allows to uniquely identify each taxi and corresponding trips as well as to study their duration and distance traveled. Only trips in the *Busy*, *Free*, *Pickup* and *CentralPickup* states are considered for this thesis as the purpose is to study traffic congestion. After dropping a passenger, taxi drivers are required to return immediately to a taxi stand but their mobility in the *Free* State (e.g. in terms of average speeds) is considerably different than in the states of *Busy* and *Pickup* (see Figure 3.8).

3.2.2 Meteorological Information

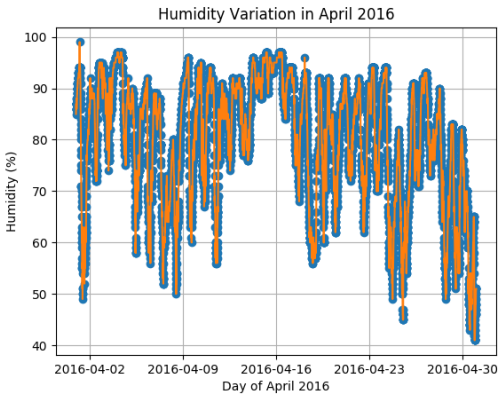
Weather states and its different features, such as, temperature or rainfall, have proven to be of great importance to traffic flow (Cools et al., 2010). Thus, weather data for the urban area of Porto was collected since it is expected to be an important factor in traffic estimation. The source for that information is *WeatherUnderground*[3] and the relevant information collected was the temperature, humidity and wind speed. All of the measures are taken with a frequency of 5 *min* and their average is later calculated according to the time sampling needed. Other features such as dew point or soil temperature were deemed irrelevant as they provided no useful information. The granularity for this dataset was of 5 *min* for the whole month of April. The expected number of measures was of 8640 (12 groups of 5 minutes per hour $\times$ 24 hours $\times$ 30 days), but due to missing data, only approximately 8100 values were available.

Figure 3.2 (a) shows the distribution of the temperature throughout the month of April. It is evident the variation due to the day/night cycle and an overall increase as the month goes on. Humidity is the amount of water vapor present in the air measured in percentage scale. In Figure 3.2 (b), the variation is displayed and it agrees with what is expected: higher humidity during the night, and lower during the day. Wind Speed refers to the speed of the wind measured in kilometers per hour. According to Figure 3.2 (c) there is no discernible pattern, as the wind speed is dependent on more features other than the day/night cycle or temperature.

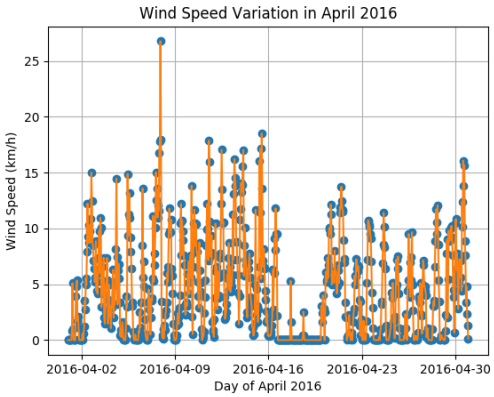
[3] https://www.wunderground.com/



(a) Temperature variation.



(b) Humidity variation.



(c) Wind speed variation.

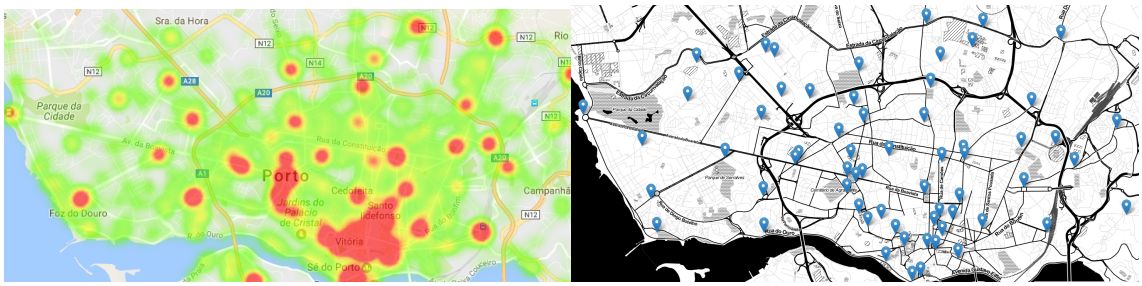
Figure 3.2: Meteorological information for the month of April, 2016.

3.3 Data Exploration

Using the data at hand, it is now possible to characterize the trips characteristics, taxi count and spread and velocity's distribution. Looking at the geographical coordinates of each trip and picking only the origin and destination, it is possible to study the start and ending positing of each trip. With that information, together with the location of the taxi stands, it is possible to see a clear correlation between the start and ending of trips and the location of said taxi stands. Figure 3.3 (a) displays the distribution of the origin and destination of the taxi trips, showing a higher concentration of services in the downtown area of the city as well as main transportation hubs, business and hotel areas. It will be later seen that the main areas of focus for this project will be the areas surrounding larger groups of taxi stands.

3.3.1 Taxi Count and Trip's Distribution

Figure 3.4 (a) shows the Cumulative Distributive Function (CDF) of the trip's duration, showing that approximately 90% of the trips have a duration lower than 750s. Figure 3.4 (b) shows that

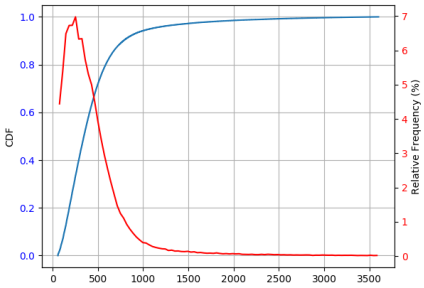


(a) Frequency of each trip's origin and destination.

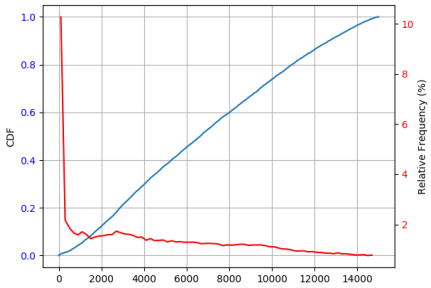
(b) Taxi stands location.

Figure 3.3: Trips and taxi stops distribution.

around 10% of the trips have a length smaller than 1000m. These small and short trips are expected as the majority of the users uses taxis for very specific travels (medical appointments too early in the morning, late night travels, etc.) where the public transportation does not cover or does not have good enough frequency.



(a) Distribution of Trip's duration and respective CDF.



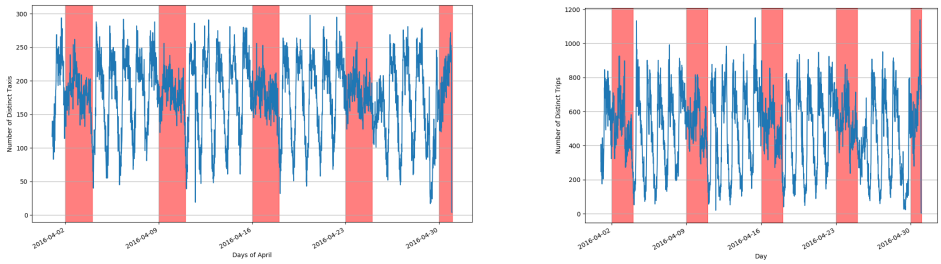
(b) Distribution of Trip's length and respective CDF.

Figure 3.4: Before and after applying filters.

Another relevant information is the distribution of the number of trips and taxis throughout the month and day, which is displayed in Figures 3.5 (a) and (b). Looking at both, it is evident the correlation between them, as the greater the number of taxis, the greater the number of trips. It is possible to see that during the week, the number of active taxis and trips fluctuates greatly (50 to 300 active taxis and 100 to 900 trips), while during the weekend, it keeps a steady number (150 to 200 taxis and 400 to 800 trips). This means that even though there are not as many active taxis, they remain with the same activity level throughout the weekend. This happens due to the great night activity that the city of Porto has, along with a younger population. There are some exceptions to the previous premises, namely day 25 and 29 of April, where the number of taxis and trips is clearly lower than expected for a week day. This happens because the April 25 is a holiday in Portugal, hence why that day behaves similarly to a weekend. On April 29 a taxi strike [4]

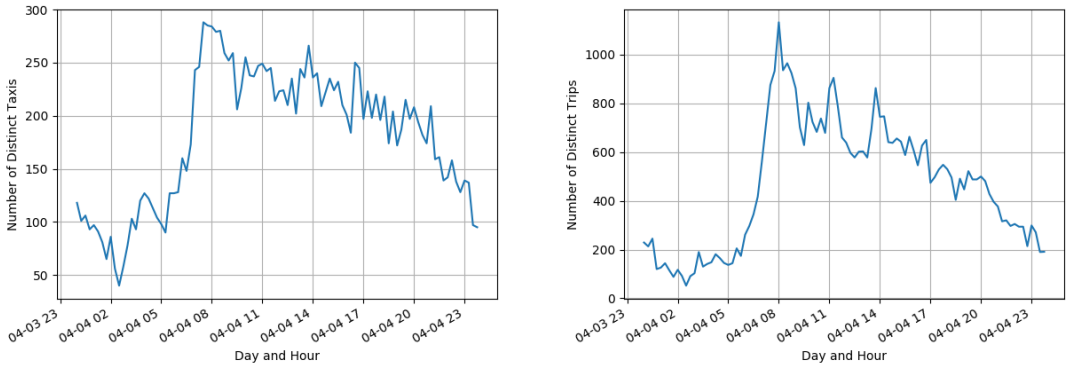
[4] http://www.dn.pt/sociedade/interior/taxistas-fazem-marcha-lenta-contra-a-uber-na-sexta-feira-5142175.html

happened and there was a very small amount of active taxi drivers, which explains the low number of trips for that day. Also, the number of trips and taxis both have daily variations possibly due to a higher number of taxi users during the day than at late night hours. That variation can be seen in further detail in Figure 3.6 where it is possible to see that at late night hours (1h-4h in the morning) the number of taxis and trips is low. Then at around 5h in the morning both start to increase until the peak is reached at around 8h in the morning. This translates to a lower amount of data in between 1h and 4h in the morning, which might decrease the amount of information retrieved for these hours. Throughout the day the number of taxis and trips remain high but with an overall decrease down to around 100 taxis and 200 trips.



(a) Distribution of Taxis per day, for the month of April. (b) Distribution of Trips per day, for the month of April.

Figure 3.5: Taxis and Trips distribution. Areas in red refer to weekends.



(a) (b)

Figure 3.6: Number of Taxis and Trips per day - 4^{th} of April.

Also relevant is the number of trips per taxi state, displayed in Figure 3.7, where it is possible to see the greater number of trips in the *Free* state when compared to the other three states. Trips in *Central Pickup* are also higher than *Pickup*, due to higher use of phone applications used to call taxis, as well as calling the taxi main central, rather than waiting at the taxi stand. Lastly, the *Busy* represents the second largest group of taxi states.

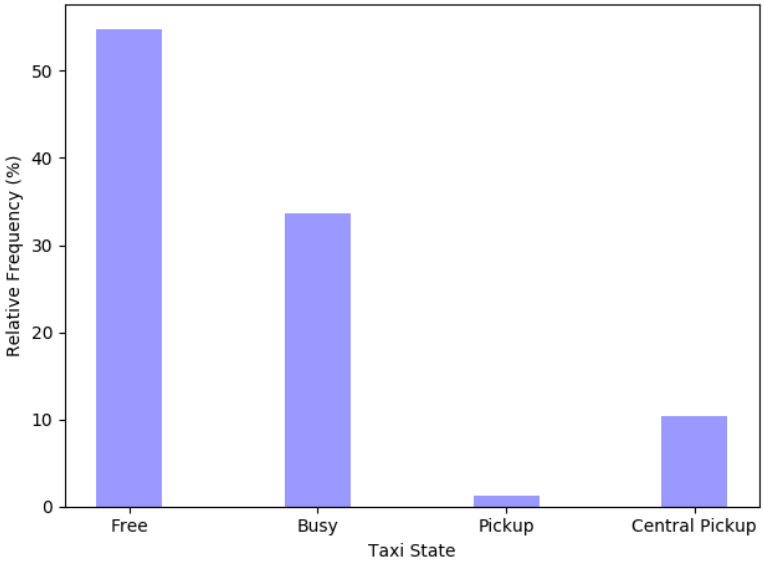


Figure 3.7: Number of trips per taxi state.

Figure 3.8 shows the CDF's for the velocities of each state in a logarithmic scale. Looking at the blue and orange lines it is possible to see that *Central Pickup* and *Free* states present the slowest speeds, while *Busy* (green) and *Pickup* (red) represent the fastest. This happens because the fastest a trip ends, the faster a taxi driver is available for the next one, increasing the overall number of trips per day. Also because some taxi companies enforce a limit to how long a user can wait for a taxi, which hastes taxi drivers to reach a pickup user.

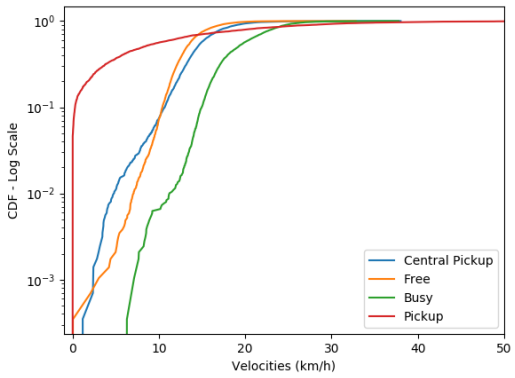


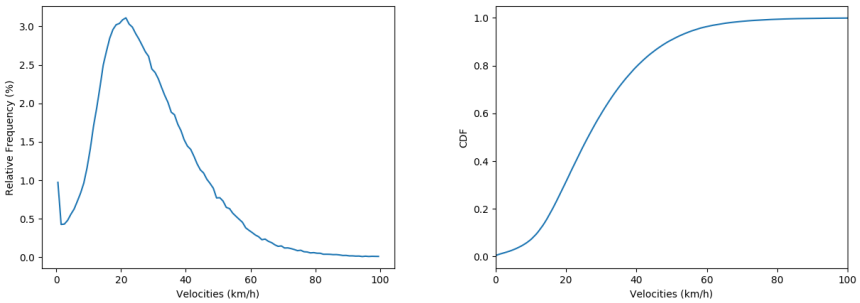
Figure 3.8: Speed CDF for each Taxi State - Logarithmic Scale.

3.3.2 Velocity's Analysis

Another very important parameter to congestion prediction besides traffic volume, is velocity and its distribution (HCM, 2000). Figure 3.9 shows a preliminary speed distribution, without any resample or aggregation. The bulk of the speeds distribution is among the lower speeds (10 to

30 km/h) and around 80% of all the velocities are lower than 40 km/h, compatible with an urban area with a great amount of traffic. Looking at the difference of speed distributions between the week and weekends in Figure 3.10, it is shown that the speed in weekends is higher than during the week, most likely due to lower amount of taxis during the weekend displayed in Figure 3.5 (a). Furthermore, the average speed decreases in the early morning (6 to 7h) reaching one of the lowest values, both in weekdays and weekends. Even though the speed increases shortly after in the day (8 to 10h), it maintains an overall low value, until it reaches its slowest value at around 17h (weekdays) and 15h (weekends). This is compatible with the higher traffic volume of the city's habitants entering (early morning) and leaving (late afternoon/evening) the metropolitan area.

Also displayed in Figure 3.10, is the average speed for the same period, when the difference between week days and weekends is not taken into account. The hours where the speed is lower than the average value - 7h, 16h, 17h and 18h - display hours with greater traffic volume, compatible with the congestioned hours of the city.



(a) (b)

Figure 3.9: Average Speed Distribution - Histogram and CDF.

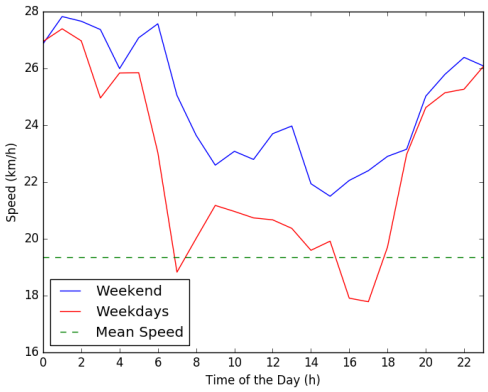


Figure 3.10: Average Speed per Hour Distribution - Weekday vs Weekend.

Chapter 4

Methodology

This chapter describes the steps taken to reach the proposed objectives. The chapter starts by explaining the preprocessing required to "clean" the data at hand. Next, a series of discretizations are applied, as well as their reasoning. Lastly, the different classification models used and developed on this project are explained.

4.1 Introduction

Figure 4.1 provides an overview of the methodology used.

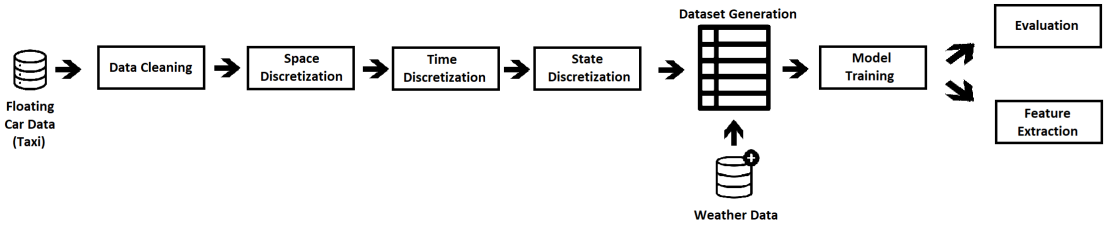


Figure 4.1: Methodology Diagram.

For Data Cleaning, different filters were applied in both the trip's duration (removed trips with a duration of less than 60 s or larger than 3600 s) and traveled distance (discarded trips with a distance less than 500m or larger than 15000m), as well as for speed outliers' removal (Hampel filter). Next, three types of discretization were applied to sort and aggregate the data at hand under a common reference grid and time division. Using weather data from external sources, and using the previously processed data, a dataset is created to train and test models in order to predict and classify the traffic's state. Lastly, different prediction models with different data aggregations are developed which are then evaluated and interpreted using different metrics for both prediction and interpretability. These steps are explained in detail in the following sections and are built with the goal to estimate traffic congestion and its propagation throughout the city of Porto, as well as to understand its most important causes.

4.2 Preprocessing

In this stage, filters were applied to remove abnormal behavior. These focused on three areas: speed value, trip duration and trip length. First, a Hampel Filter (Pearson et al., 2016) is applied to the speed distribution in order to remove or reduce the impact of outliers due to GPS errors. Figure 4.2 shows a profile speed example before and after applying the Hampel filter. As expected, several outliers are removed (marked in red), caused by erroneous GPS positions, that affect the speed calculation, resulting in impossible or improbable values. Hampel is a standard median filter based on a symmetric moving window. It replaces the central value in the data window with the median if it lies far enough from the median to be deemed an outlier.

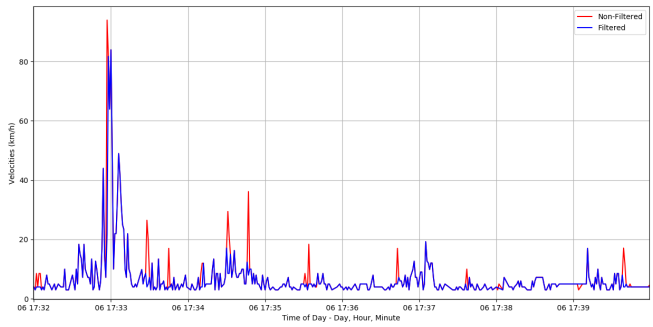


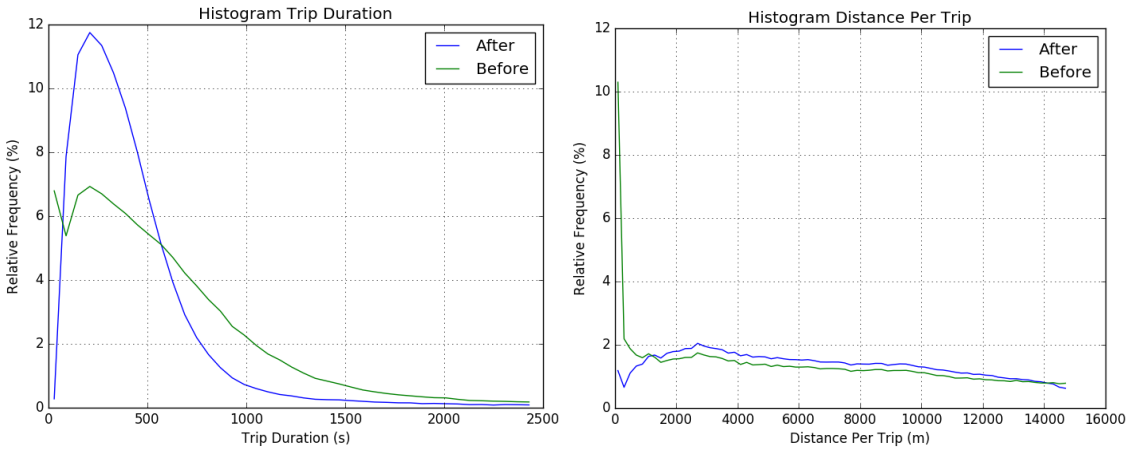
Figure 4.2: Comparison between before and after applying an Hampel Filter.

Next, two other filters were applied, now to the trip's duration and length. First, trips with a duration smaller than 60 seconds or larger than 3600 seconds were removed (Figure 4.3 (a)). This resulted in the removal of all the null-duration trips and trips too long, increasing the relative frequency of the smaller trips. The last filter was applied to the trip's length and it removed all the trips with a length less than 500 meters or higher than 15000 meters (Figure 4.3 (b)). This resulted in the removal of approximately 9% of the total trips. Trips with a large length were removed because they most likely went outside the metropolitan area of the city, which is the focus of this project. In total, around 110.000 trips were removed, lowering the value from almost 400 thousand to approximately 280 thousand.

4.3 Discretization

4.3.1 Space Discretization

Following the approach by Castro et al. (2013a), the city of Porto was divided in a grid of equally sized square cells with 250 (Figure 4.4), 500 (Figure 4.5) or 750 (Figure 4.6) meters of edge, covering over $63km^2$. This allows to study the city in greater detail, focusing in a cell at a time, instead of the full road network.



(a) (b)

Figure 4.3: Comparison of trips duration (a) and length (b) before and after applying filters.

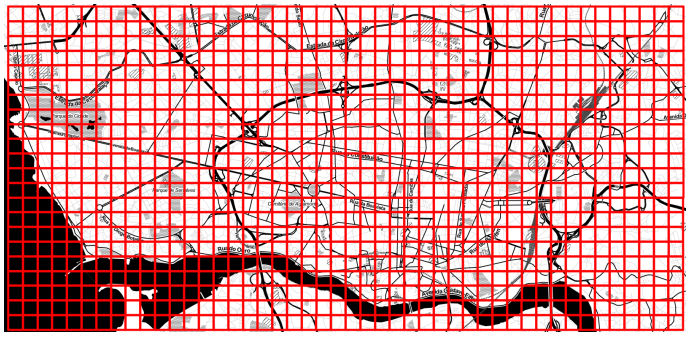


Figure 4.4: Grid Size 250m.

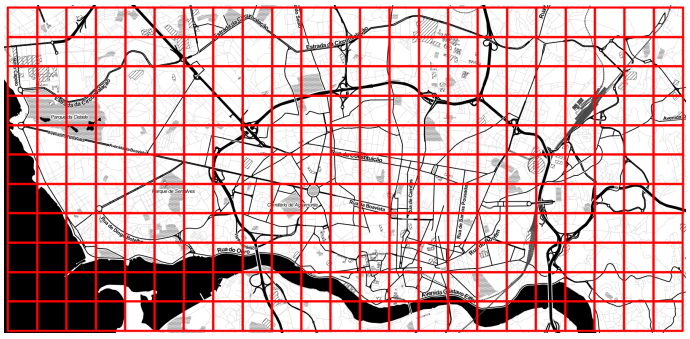


Figure 4.5: Grid Size 500m.

Looking at Figure 4.7, a comparison of the grid sizes is displayed. The metrics used are the speed values, number of data values per slot of 15 minutes (see section 4.3.2) and the speed standard deviation.

In Figure 4.7 (a) all grid sizes have a considerable amount of slots with very few speed measures, with the 750*m* grid size displaying a higher amount of time slots with more than 100 speed

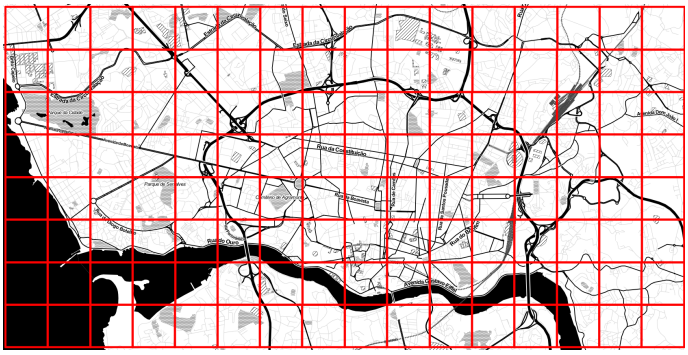
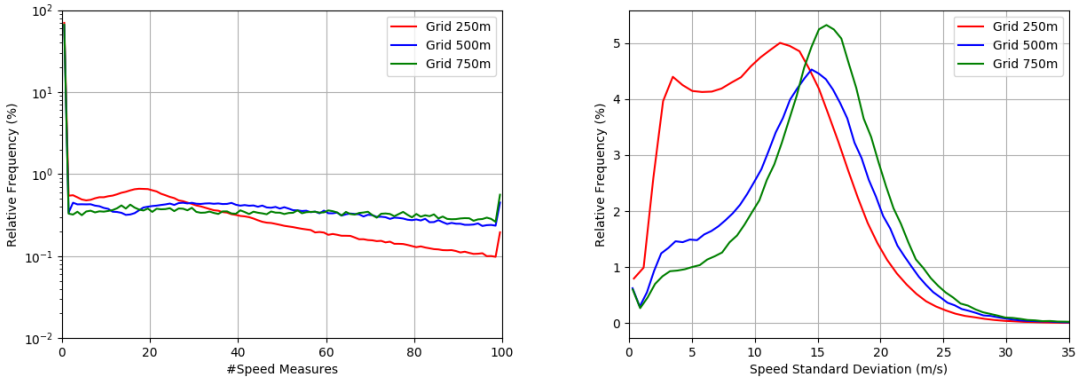


Figure 4.6: Grid Size 750m.

measures. This is explained by greater cell size providing more information per cell.

In Figure 4.7 (b), grid size 250m displays a lower standard deviation, albeit with a broader range. Regardless, grid 500m and 750m all display higher standard deviations because of the higher amount of information per cell, which is prone to greater errors in the speed measurement.

The selection of grid size was made considering the trade-off between data availability (number of speed measures) and variability in traffic state estimation (standard deviation). Even though larger cells improve data availability, this increase in the size of the covered area per cell will lead to increased variability of the measurements. Another problem with larger cells is the increase in the probability of co-existence of different traffic states in a given area due to the larger road length and number of intersections (shown in the next section). Since there isn't yet an immediate better grid size, all three sizes will be used and analyzed in order to determine after the results, the most adequate size for the goals of this project.



(a) (b)

Figure 4.7: Speed Measures and Standard Deviation per Grid Size.

4.3.1.1 Cell Selection and Analysis

In order to reduce the impacts of data sparsity and to be able to focus on the urban area, ten cells were selected for each grid size. The criteria used was the amount of information. This would not only ensure a solid quantity of information, but it also reduces the processing time and sheer amount of data. The ten cells with most information are presented in Figure 4.8 for different grid sizes. As expected, many of the cells with the highest amount of information are located in the downtown area, which are also the zones with highest count of taxi stands (See tables 4.1, 4.2, and 4.3, row Taxi Stands). Other zones with a great amount of information out of downtown are main avenues with lots of traffic (Figure 4.8 (b), cells 6,10 and 7,10 for example). Also, due to the finer granularity, some cells of grid size 250 are located not only in the downtown area, but also in one of the main train stations of Porto, as well as in a hospital (cells 16,35 and 1,30 of Figure 4.8 (a), respectively).

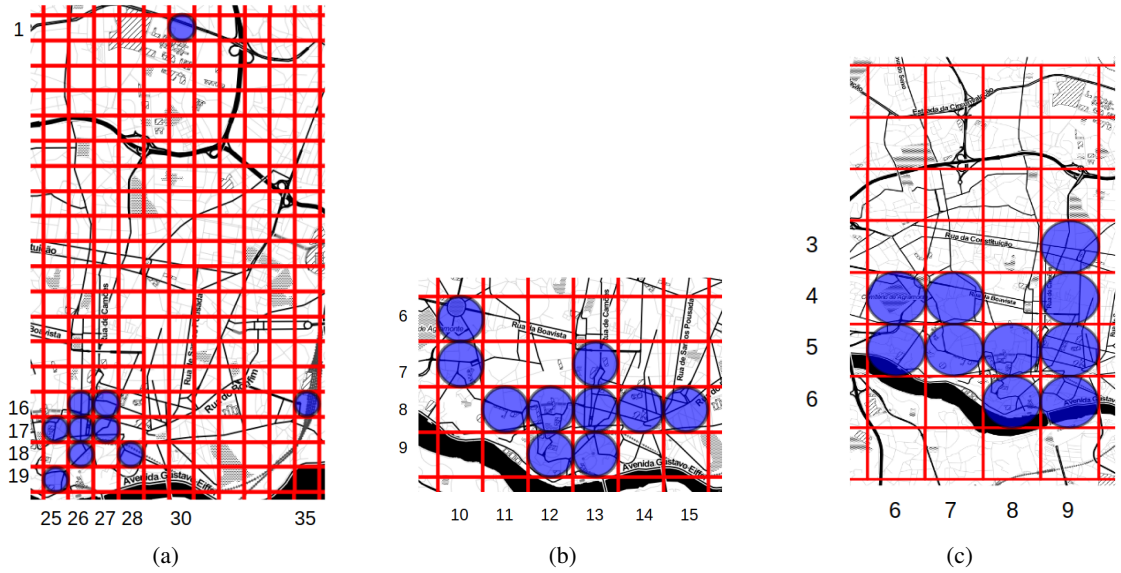


Figure 4.8: Top 10 Cells for grid sizes 250 (a), 500 (b) and 750 (c).

A better characterization of each cell is required, in terms of graph analysis to describe the number of streets, their length, network connectivity and other useful metrics which might have an impact on the results. In order to achieve that, the *OSMNX* tool was used (Boeing, 2017) which is capable of analyzing the street networks of the different sizes. The characterization of the top 10 cells is presented in tables 4.1, 4.2 and 4.3. *OSMNX* is a package for *Python* that does graph analysis using information from *OpenStreetMap* [1]. With it, it's possible to retrieve important statistics (Boeing, 2017) for the full city as well as a comparison within the top 10 cells selected for each grid size. The metrics used to evaluate each cell are:

- Number of Nodes - Total number of nodes in each section. Nodes are the basic elements of graph theory, and in this analysis represent the end or beginning of a street;

[1] https://www.openstreetmap.org/

- Node Density - Number of nodes divided by the area (in km^2) of the section at study;
- Average Node Degree - Average number of edges connected to nodes;
- Number of Intersections - Total number of street intersections in each section. In this analysis, the number of intersections is the same as the node's with more than one edge connected to it;
- Intersection Density - Number of intersections divided by the area of the section at study (per km^2);
- Number of Edges - Total number of edges in each section. Edges are the connection between nodes;
- Total Edge Length - Sum of edge lengths in the section at study (m);
- Edge Density - Total edge length divided by the area of the section at study (km/km^2);
- Average Edge Length - Average edge length in the section at study (m);
- Number of Streets per node - Number of streets emanating from the nodes and the respective count for that number;
- Average Streets per node - Average number of streets that emanate from each node;
- Total Street Length - Sum of all street's length;
- Street Density - Total street length divided by the area of the section at study (km/km^2);
- Average Street Length - Average street length in the section at study (m);
- Number of Taxi Stands - Total number of taxi stands in the section at study;
- Number of Traffic Lights - Total number of traffic lights in the section at study;
- Eccentricity - represents the maximum distance from each node, to all others nodes, weighted by length (m);
  - Diameter - Maximum eccentricity of any node (m);
  - Radius - Minimum eccentricity of any node (m);
- Average Node Connectivity - Expected number of nodes that must be removed to disconnect randomly selected pair of non-adjacent nodes. Connectivity measures the resilience of a section, as complex networks with high connectivity provide more routing choices and are more robust against congestion or other kinds of failure;
- Node Connectivity - Minimum number of nodes that must be removed to disconnect the section at study;
- Edge Connectivity - Minimum number of edges that must be removed to disconnect the section at study;
- Betweenness Centrality - The fraction, for each node, of all shortest paths that pass through the node. The highest possible value represents the proportion of shortest paths that pass

through the most important node/edge. This is another indicator of resilience, as sections with a high maximum betweenness centrality are more prone to failure should this single choke point fail;

  - Average Betweenness Centrality - Average value of all the Betweenness Centralities of all the nodes of the section at study;

- PageRank - Ranking of nodes based on the structure of incoming edges;
- Minimum PageRank - Lowest PageRank of any node in the section at study;
- Maximum PageRank - Highest PageRank of any node in the section at study;

While looking at the tables there are some missing values for the full grid, that could not be calculated either due to performance constraints (total grid values) or as a result of some cells presenting a graph with little information such as only one node or one-way streets, to calculate some of the metrics (grid 250m, cells 16,26, 17,27, 18,28 and 1,30). But the remaining data shows that the greater the size of a cell, the greater overall amount of nodes, edges, intersections, streets per node, edge length, street length and its average, number of taxi stands and traffic lights, just as expected as greater cells have more nodes, edges and streets. The average node degree, diameter and radius also increase with the greater cell sizes. This happens because as the cell increases, the number of edges does too, and so, the more edges are inbound and outbound of each node, increasing its average degree. The diameter and radius increase because they are dependent of the number of nodes and the distances between them. Greater cells can have a larger space between node when compared to smaller cells, increasing the eccentricity overall.

On the other hand, node density, edge density and average betweenness centrality increase as the grid size decreases. This happens because even though the cells are smaller, they focus on the main zones of traffic, like downtown. On these zones the number of nodes is similar and when divided by a greater area due to the size of the cell, decrease their overall value.

Lastly, edge density, average street per node, street density, average node connectivity, node and edge connectivity, and minimum and maximum pagerank all stay roughly the same with as the grid sizes changes. The reason for this is due to all cells of the different grid sizes tend to focus on the same zones of traffic (overall, there are exceptions), and so the edge density, average streets per node and street density stay almost the same, even though the respective total edge length, number of streets per node and number of streets decrease, as the area also decreases. As for Node and Edge Connectivity, these are always 1, due to presence of dead ends. Should that node or edge be removed and it would automatically disconnect the network - there would be no way to reach the node or edge that remains. The Average Node Connectivity also stays roughly the same on all cells and in low numbers, showing that each cell is not very resilient. This happens because they are part of a network greater than them. Lastly, PageRank stays overall the same for grid sizes, again, because the cells tend to focus on the same areas with more traffic.

Network Measures	**Unit**	**Total Grid**	**Cell 17,26**	**Cell 17,25**	**Cell 16,26**	**Cell 19,25**	**Cell 17,27**	**Cell 18,26**	**Cell 18,28**	**Cell 16,35**	**Cell 16,27**	**Cell 1,30**
Cell's Coordinates (Center)	Degrees (°)		41.14659, -8.61217	41.14659, -8.61516	41.14659, -8.61217	41.14209, -8.61516	41.14659, -8.60919	41.14434, -8.61218	41.14434, -8.60620	41.14884, -8.58530	41..14883, -8.60918	41.18256, -8.60018
Number of Nodes		7031	15	17	11	16	6	13	3	5	8	4
Node Density	*node per km²*	111.2	240	272	176	256	96	208	48	80	128	64
Average Node Degree		4.35	2.27	2.71	2.18	3.25	1.67	3.38	1.33	2.00	3.25	2.00
Number of Intersections		6153	15	17	11	16	6	12	6	5	8	4
Intersection Density	*intersection per km²*	97.3	240	272	176	256	96	192	48	80	128	64
Number of Edges		15298	17	23	12	26	5	22	2	5	13	4
Edge Density	*km/km²*	20864	17322.12	18086.34	9458.72	15558.87	5543.72	14238.75	4309.22	1651.96	15230.43	3583.76
Total edge length	*meters*	1319650	1054.48	1130.40	591.17	972.43	346.48	889.92	269.33	103.25	951.90	223.99
Average Edge Length	*meters*	86.26	62.03	49.15	49.26	37.40	69.30	40.45	134.66	20.65	73.22	56.00
Number of Streets per node		0: 0 1: 878 2: 58 3: 5260 4: 790 5: 42 6: 3	0: 0 1: 0 2: 0 3: 13 4: 2	0: 0 1: 0 2: 1 3: 15 4: 1	0: 0 1: 0 2: 0 3: 8 4: 3	0: 0 1: 0 2: 1 3: 13 4: 2	0: 0 1: 0 2: 0 3: 5 4: 1	0: 0 1: 1 2: 4 3: 7 4: 1	0: 0 1: 0 2: 0 3: 1 4: 2	0: 0 1: 0 2: 0 3: 4 4: 1	0: 0 1: 0 2: 1 3: 3 4: 4	0: 0 1: 0 2: 0 3: 4
Average Streets per Node		2.87	3.13	3.0	3.27	3.06	3.17	2.62	3.67	3.20	3.38	3.00
Total street length	*meters*	887822	1054.48	1007.20	591.17	746.23	346.48	507.92	269.33	103.25	672.23	223.99
Street Density	*km/km²*	14036.7	16871.64	16115.16	9458.72	11939.60	5543.72	8126.69	4309.22	1651.96	10755.64	3583.76
Average Street Length	*meters*	88.54	62.03	47.96	49.26	39.28	69.30	42.33	134.66	20.65	74.69	56.00
Number of Taxi Stands		63	2	0	1	1	0	0	1	1	1	1
Number of Traffic Lights		290	2	4	2	1	1	1	3	1	1	1
Diameter		-	185.34	381.34	*	451.03	*	253.12	*	72.87	118.51	*
Radius		-	138.39	242.61	*	236.16	*	148.75	*	63.57	59.72	*
Average Node Connectivity		-	0.48	0.60	0.45	0.95	0.40	0.72	0.50	0.80	0.88	0.58
Node Connectivity		-	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Edge Connectivity		-	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Average Betweenness Centrality		-	0.09	0.11	0.08	0.24	0.09	0.14	0.17	0.30	0.09	0.08
Minimum PageRank		-	0.018	0.018	0.026	0.011	0.067	0.012	0.184	0.030	0.019	0.128
Maximum PageRank		-	0.154	0.108	0.165	0.203	0.298	0.227	0.474	0.253	0.362	0.424

Table 4.1: Grid Analysis - Full Grid 250m vs. Top 10 Cells.

Network Measures	**Unit**	**Total Grid**	**Cell 8,13**	**Cell 8,12**	**Cell 8,14**	**Cell 6,10**	**Cell 7,13**	**Cell 8,11**	**Cell 9,12**	**Cell 9,13**	**Cell 8,15**	**Cell 7,10**
Cell's Coordinates (Center)	Degrees(°)		41.14771, -8.61068	41.14771, -8.61665	41.14771, -8.60471	41.15671, -8.62858	41.15221, -8.61067	41.14771, -8.62262	41.14322, -8.61665	41.14322, -8.61068	41.14771, -8.59873	41.15221, -8.62859
Number of Nodes		7031	42	37	27	45	32	20	48	41	41	27
Node Density	*node per km²*	111.2	168.0	148.0	108.0	180.0	128.0	80.0	192.0	164.0	164.0	108.0
Average Node Degree		4.35	3.05	2.92	2.81	3.20	2.75	2.90	3.29	3.66	3.41	3.56
Number of Intersections		6153	40	36	27	44	32	20	47	34	41	23
Intersection Density	*intersection per km²*	97.3	160.0	144.0	108.0	176.0	128.0	80.0	188.0	136.0	164.0	92.0
Number of Edges		15298	64	54	38	72	44	29	79	75	70	48
Edge Density	*km/km²*	20864	17322.12	13152.65	14852.99	15823.68	15312.66	11193.49	19059.38	18403.89	18888.61	16946.83
Total edge length	*meters*	1319650	4330.53	3288.16	3713.25	3955.92	3828.17	2798.37	4764.84	4600.97	4722.15	4236.71
Average Edge Length	*meters*	86.26	67.66	60.89	97.72	54.94	87.00	96.50	60.31	61.35	67.46	88.26
Number of Streets per node		0: 0 1: 878 2: 58 3: 5260 4: 790 5: 42 6: 3	0: 0 1: 2 2: 1 3: 29 4: 10	0: 0 1: 1 2:1 3:31 4: 4	0: 0 1: 0 2: 0 3: 16 4: 11	0: 0 1: 1 2: 0 3: 41 4: 3	0: 0 1: 0 2: 0 3: 26 4: 6	0: 0 1: 0 2: 0 3: 15 4: 5	0: 0 1: 1 2: 2 3: 41 4: 4	0: 0 1: 7 2: 4 3: 27 4: 3	0: 0 1: 0 2: 0 3: 31 4: 8 5: 2	0: 0 1: 4 2: 1 3: 16 4: 4 5: 2
Average Streets per Node		2.87	3.12	3.03	3.41	3.02	3.19	3.25	3.00	2.63	3.29	2.96
Total street length	*meters*	887822	3739.30	2972.36	3283.72	3237.86	3438.23	2147.85	3927.59	3282.77	3842.59	3043.02
Street Density	*km/km²*	14036.7	14957.20	11889.45	13134.89	12951.42	13757.00	8591.39	15710.35	13136.65	15370.36	12172.06
Average Street Length	*meters*	88.54	66.77	60.66	93.82	54.88	83.88	89.49	61.37	67.02	67.41	89.50
Number of Taxi Stands		63	4	2	0	4	1	0	1	0	1	1
Number of Traffic Lights		290	6	8	11	2	6	6	2	4	7	5
Diameter		-	993.95	381.34	1044.60	1374.20	174.99	495.04	1207.85	734.31	1153.76	881.93
Radius		-	547.52	242.61	638.93	578.49	98.48	268.76	528.39	341.83	614.96	451.51
Average Node Connectivity		-	0.85	0.43	0.81	0.84	0.37	0.53	0.78	0.82	0.95	0.77
Node Connectivity		-	1	1	1	1	1	1	1	1	1	1
Edge Connectivity		-	1	1	1	1	1	1	1	1	1	1
Average Betweenness Centrality		-	0.11	0.05	0.11	0.17	0.04	0.08	0.11	0.11	0.11	0.10
Minimum PageRank		-	0.004	0.007	0.006	0.004	0.009	0.010	0.004	0.004	0.005	0.006
Maximum PageRank		-	0.083	0.066	0.148	0.065	0.068	0.214	0.076	0.058	0.044	0.129

Table 4.2: Grid Analysis - Full Grid 500m vs. Top 10 Cells.

Network Measures	**Unit**	**Total Grid**	**Cell 5,8**	**Cell 5,9**	**Cell 6,8**	**Cell 5,7**	**Cell 5,6**	**Cell 4,9**	**Cell 4,7**	**Cell 6,9**	**Cell 4,6**	**Cell 3,9**
Cell's Coordinates (Center)	Degrees(°)		41.14884, -8.61515	41.14884, -8.60620	41.14209, -8.61516	41.14884, -8.62411	41.14884, -8.63307	41.15558, -8.60619	41.15558, -8.62410	41.14209, -8.60620	41.15558, -8.63306	41.16233, -8.60618
Number of Nodes		7031	73	66	79	37	52	73	64	37	63	40
Node Density	*node per km²*	111.2	129.78	117.33	140.44	65.78	96.0	129.78	113.78	65.78	112.00	71.11
Average Node Degree		4.35	2.93	3.33	3.60	2.97	4.00	3.34	3.50	3.57	3.46	3.00
Number of Intersections		6153	70	61	75	35	42	67	60	29	59	38
Intersection Density	*intersection per km²*	97.3	124.44	108.44	133.33	62.22	78.22	119.11	106.67	51.56	104.89	67.56
Number of Edges		15298	107	110	142	55	102	122	112	66	109	60
Edge Density	*km/km²*	20864	13233.70	17852.01	16974.88	10578.86	16976.05	16526.95	15642.49	9809.24	16326.65	11022.40
Total edge length	*meters*	1319650	7443.95	10041.76	9548.37	5950.61	9055.77	9296.41	8798.90	5517.70	9183.74	6199.84
Average Edge Length	*meters*	86.26	69.57	91.29	67.24	108.19	88.42	76.20	78.56	83.60	84.25	103.34
Number of Streets per node		0: 0 1: 878 2: 58 3: 5260 4: 790 5: 42 6: 3	0: 0 1: 3 2: 3 3: 57 4: 10	0: 0 1: 5 2: 1 3: 36 4: 24	0: 0 1: 4 2: 7 3: 61 4: 7	0: 0 1: 2 2: 1 3: 26 4: 8	0: 0 1: 10 2: 0 3: 40 4: 1 5: 1	0: 0 1: 6 2: 0 3: 50 4: 17	0: 0 1: 4 2: 0 3: 49 4: 11	0: 0 1: 8 2: 0 3: 26 4: 3	0: 0 1: 4 2: 0 3: 47 4: 9 5: 3	0: 0 1: 2 2: 0 3: 24 4: 12 5: 2
Average Streets per Node		2.87	3.01	3.20	2.90	3.08	2.69	3.07	3.05	2.65	3.11	3.30
Total street length	*meters*	887822	6805.97	8254.46	7300.01	4714.11	5117.34	7399.83	6540.08	3580.38	6522.65	5732.83
Street Density	*km/km²*	14036.7	12099.50	14674.60	12977.80	8380.63	9535.95	13155.26	11626.82	6365.13	11595.83	10192.15
Average Street Length	*meters*	88.54	69.45	91.72	68.87	102.48	85.14	75.51	73.48	83.26	77.65	108.17
Number of Taxi Stands		63	5	3	2	3	0	1	2	1	2	3
Number of Traffic Lights		290	12	19	4	8	4	12	8	8	5	12
Diameter		-	1309.74	1253.22	1598.13	1259.33	1591.75	1652.93	1961.59	1348.19	2220.93	1896.10
Radius		-	460.41	693.00	732.35	774.60	808.64	970.47	731.17	858.89	988.04	913.37
Average Node Connectivity		-	0.51	0.75	0.76	0.64	0.90	0.87	1.03	1.00	0.84	0.69
Node Connectivity		-	1	1	1	1	1	1	1	1	1	1
Edge Connectivity		-	1	1	1	1	1	1	1	1	1	1
Average Betweenness Centrality		-	0.05	0.05	0.08	0.09	0.11	0.08	0.12	0.14	0.09	0.11
Minimum PageRank		-	0.002	0.003	0.002	0.005	0.003	0.002	0.003	0.006	0.003	0.006
Maximum PageRank		-	0.054	0.044	0.046	0.118	0.048	0.032	0.047	0.078	0.047	0.074

Table 4.3: Grid Analysis - Full Grid 750m vs. Top 10 Cells.

4.3.2 Time Discretization

The second discretization ensures sufficient data availability by re-sampling the time into 15 minutes intervals from the original 1Hz frequency. This, together with the previous space discretization, ensures aggregated speed (or other) measurements per cell and per 15 minutes interval. Other authors usually use the same sampling (e.g. Mehta and Chana (2017) and Zhu et al. (2013)).

When using greater sampling intervals, the amount of data is reduced, which can be useful in terms of processing time. Though it can also be hurtful in the accuracy of the value as the more you aggregate to larger bins, the less accurate the value can become. This happens due to the aggregation on the same time slot of different flow states and very different speeds from different road types. When using lower sampling intervals, the big drawback is the amount of information overall and respective large processing time, while for each slot the amount of information is reduced, reducing its importance.

With this aggregation a method to calculate the velocity is needed. The method developed, seen in Figure 4.9, first groups each taxi's velocities for a time slot, and calculates an average for each taxi. With these new values, a new average is calculated which ends being the velocity value for a time slot, and the method is repeated for all time slots with available information.

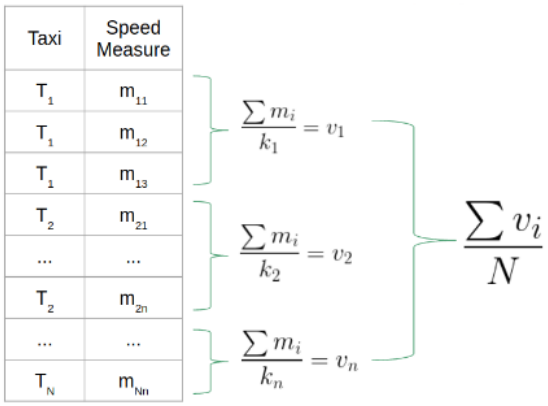


Figure 4.9: Speed Calculation for each time slot.

4.3.3 State Discretization

State discretization is required to separate the speed distribution of taxis into three levels of service (LOS): *Free Flow*, *Synchronized Flow* and *Congestion*. As represented in Figure 4.10, the three traffic states are defined as follows for urban areas (urban streets of class IV in HCM (2000)):

- *Free Flow* refers to the fastest state where each car is completely or reasonably unimpeded in their ability to maneuver within the traffic stream. In this case, the speeds typically range between 32-41 km/h [2]. Corresponds to LOS A and B in HCM (2000).

[2] Values from HCM (2000)

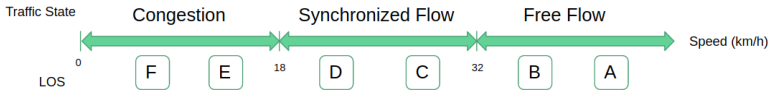


Figure 4.10: Comparison between Traffic States and Level of Service.

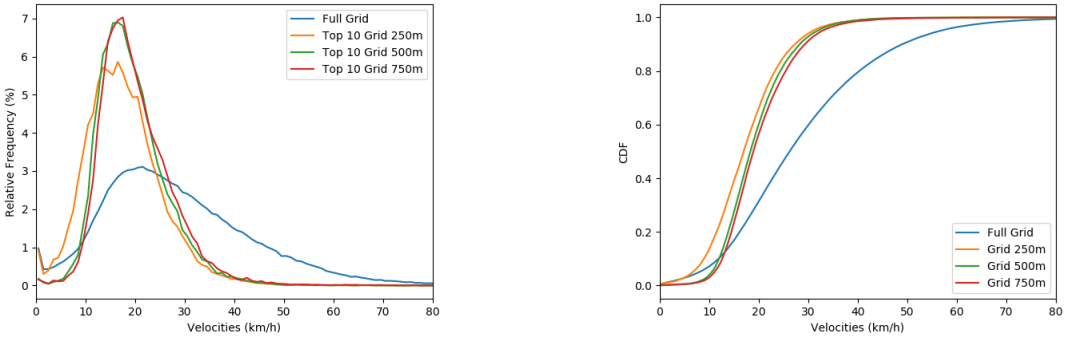
- *Synchronized Flow* represents stable operations with an ability to maneuver and change lanes restricted. Also, small increases in flow may cause substantial increases in delay and decreases in speed. Speeds in this state range between 18-23 km/h [2]. Coincides with LOS C and D in HCM (2000).

- *Congestion* is the slowest state caused by either high signal density, high volume of traffic or extensive queuing. Usual speeds of this state range from close to zero values in the most severe cases, up to 18 km/h [2]. Matches LOS E and F in HCM (2000).

Further analysis of speed distribution was performed which allowed to refine the limits of speed for each level. This analysis was performed on the 10 cells - Top 10 - with the highest amount of information - higher number of taxis and trips in the selected month. Figures 4.11 (a) and (b), show the apparent higher number of taxis with speeds between around 10 km/h and 30 km/h for grid sizes 500m and 750m, while grid 250m is shifted to slightly lower values, from 5m to 30km/h. These values limit the bulk of the speed distributions for the different grid sizes and peak at 18km/h for grids 250m and 500m, while grid 750m peaks at 20km/h. When compared with the full grid, it shows a reduction of velocities higher than 35km/h (less than 0.5% when compared to 2%) as expected from the top 10 cells, since they focus on urban areas with a lot more traffic, increasing congestion. Figure 4.12 represents the cumulative distribution function (CDF) of the speeds for each of the top 10 cells in the grid size of 500m, which displays cells with higher number of slower speeds (e.g. cell 8_13 and 8_12), while others are faster cells (e.g. 7_13 and 7_12). Overall, the speeds between the different cells tend to shift right, increasing the number of higher speeds. Looking at cells 8_12 and 8_13, the ones on the far left, it is understandable why they have a high amount of lower speeds, as they are located in the heart of the urban area, where the congestion reaches its peak either by the amount of traffic or the high amount of traffic lights. On the right, with higher velocities, are cells 7_10 and 7_13, which can be explained due to the lower amount of nodes (27 and 32 respectively) present in those roads, even though they are fairly close to the center of the urban area.

Analyzing the relationship between the average speed of the top 10 cells for each grid and their standard deviation in Figure 4.13, it shows a higher standard deviation for grid size 250m (first mentioned in 4.3.1.1) but overall the three grid sizes have a very similar correlation coefficient between the speeds and their standard deviation.

Next, Figures 4.14 represent Macroscopic Fundamental Diagrams (MFD) of average speed of taxis per time slot of 15 minutes versus the number of unique taxis that exist in the same slot. The Figures displayed have no aggregation and are for all grid sizes. MFD is a type of graphic that represents the traffic flow by relating the amount of different taxis, their density and speeds. For



(a) (b)

Figure 4.11: Speed Distribution Comparison - Histogram and CDF

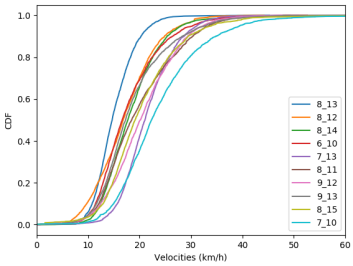


Figure 4.12: CDF of Top 10 Cells for grid size 500m.

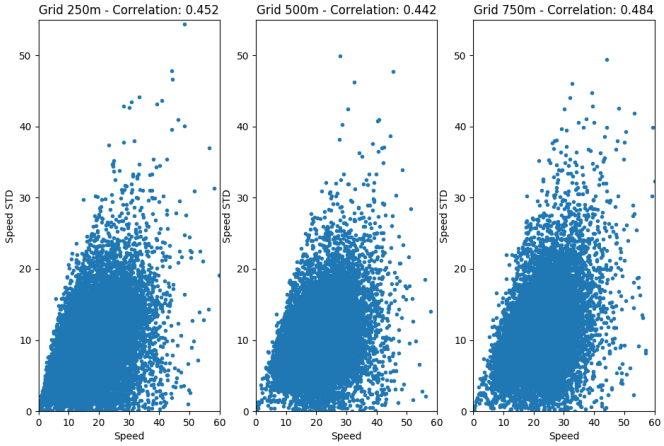
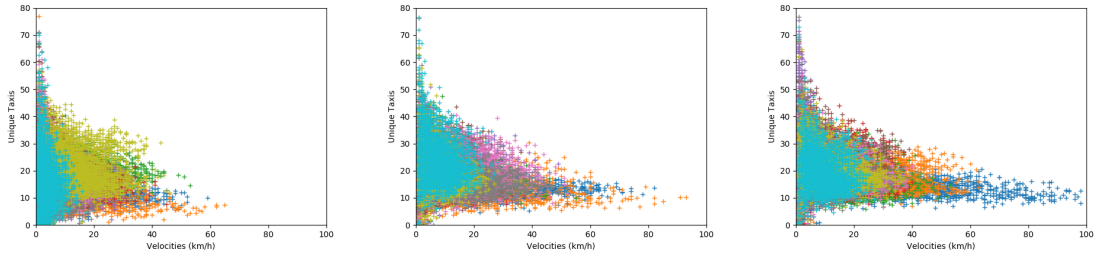


Figure 4.13: Correlation and respective coefficients between average speed and standard deviation.

this instance, in the x-axis, the number of different taxis per time slot of 15 minutes is calculated by considering that every time a taxi enters a cell, it counts as a new taxi. For the y-axis, the velocities of each taxi are calculated, by averaging for each taxi, in each time slot, as explained in the section 4.3.2.

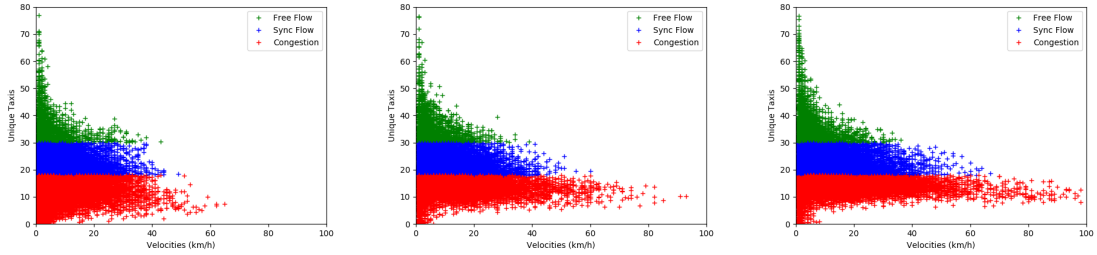
These Figures show that as the grid size increases, the MFD's shape gets closer to a multiplicative inverse curve and values of speed increase. This happens because each cell has more information in the greater grid sizes. Also shown on all grid sizes, is the apparent existence of a higher value in the taxi count for the velocities between roughly $10km/h$ and $20km/h$. Using this information, and the speed limits for the different LOS mentioned before, the limits to separate each traffic state were chosen as 18 km/h to separate between *Congestion* and *Synchronized Flow*, and 30 km/h to differentiate amid *Synchronized Flow* and *Free Flow*.

Through these values, all 2880 slots of 15 minutes of the month of April (4 slots per hour $\times$ 24 hours per day $\times$ 30 days) were classified in one of the three mentioned LOS (See Figure 4.15 (b)) for all the top 10 cells. This will be known in the further steps as the ground truth for the prediction work.



(a) (b) (c)

Figure 4.14: MFD Comparison for all grid sizes.



(a) (b) (c)

Figure 4.15: MFD Comparison and Discretization for all grid sizes.

This discretization resulted in the traffic state distribution seen in Figure 4.16 where it is discernible the small amount of time slots in the state of *Free Flow* and the large amount in *Congestion*. This is compatible with the reality scenario as the top 10 cells are mainly located in urban areas under heavy, where the velocities do not usually get much higher than $30km/h$.

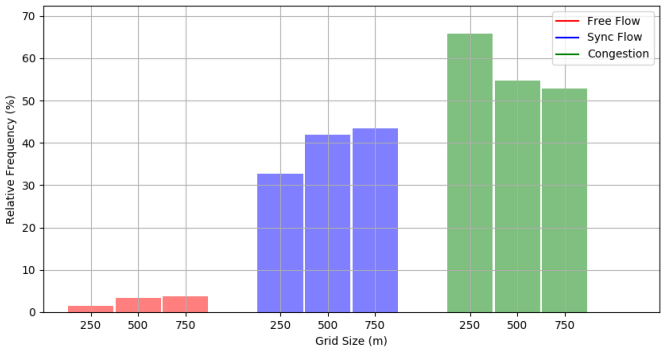


Figure 4.16: Mass Distribution Function for the Traffic States after State Discretization.

4.4 Dataset Generation

After classifying the traffic states, tables were created that appended more information for the same time slot and for a specific cell. The list of features for each slot of time and each cell are:

- Cell Center Coordinates - Latitude and Longitude
- Hour = {1, 2, 3, ..., 24}
- Minutes (time slot of 15 *min*)
- Day of the month = {1, 2, 3, ..., 30}
- Weekday = {Monday, Tuesday, Wednesday, Thursday, Friday}
- Weekend = {Saturday, Sunday}
- Average Temperature (^oC)
- Average Wind Speed (km/h)
- Humidity (%)
- LOS of the previous 4 time slots (1 hour window) for that cell as well as the top 10 cells (See Figure 4.17)
- Acceleration and its statistics of a cell including its previous states as well as the values for that time slot:
  - Quartile 25%
  - Median
  - Quartile 75%
  - Standard Deviation
- Following Speed statistics of a cell including its previous states as well as the values for that time slot:
  - Quartile 25%

- Median
- Quartile 75%
- Standard Deviation
- Interquartile-Range (IQR)

Using these features, the target is to predict the future LOS of a cell, classifying in one of the three mentioned traffic states.

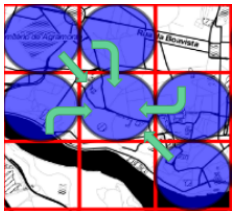


Figure 4.17: Representation of using other cells' previous states. Each arrow represents information from either 15, 30, 45 or 60 minutes of previous information.

4.5 Data Preparation

Before the data set could be used for model training, it had to prepared in order to reach better results, as well as to use the maximum amount of data. The techniques used were:

- One-Hot Encoder (OHE) - From a machine learning perspective, it doesn't make sense for some features to have nominal values that could induce that one variable is smaller than the other (e.g. Monday, Day 0 and Tuesday, Day 1). For this reason, the features are replace by boolean variables on new columns;
- Categorical Variable Conversion - in order to ensure that some features were treated as categorical instead of numerical, they are converted prior to being used in the different models. The conversion was done for all the binary features, specially the ones created with OHE;
- Data Imputation - Either because of data sparsity or low taxi coverage, data for some features (mention which ones) in some specific time slots was missing for this data set. In order to recover the rest of the information, the missing data was imputed, using the mean value for that feature. In total, 12.5% of data was corrected for the acceleration features, specifically for grid size 250m.

4.6 Classification

In this section it is explained how the classification of each slot is performed. It explains how the models were created, how data was split and the evaluation methods. The performance metrics are discussed in the next sections.

4.6.1 Classifiers

The Classifiers used in this project were: Classification And Regression Trees (CART) (Breiman et al., 1984) and Random Forest (RF) (Breiman, 2001). The first was used for all models, as this project purpose is to focus on interpretability and thanks to the generated decision trees, a lot of information can be extracted. The chosen function to measure the quality of a split is *gini impurity* which is an appraise of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset (Mingers, 1989). Random forest's models were used as basis of comparison to find the best possible result solely based in terms of predictability.

4.6.2 Parameter Tunning and Model Evaluation

For each model certain hyper-parameters were tuned in order to achieve the best possible result. The parameters for CART were:

- Minimum Samples per Leaf - sets a minimum number of samples required to be at a leaf node;
  - Values used = {3, 5, 7, 9, 10, 15, 20, 50}
- Max Depth - limits the length of the longest path from the root to a leaf;
  - Values used = {3, 5, 7, 9, 11, 13}
- Class Weight - sets the weight for each class. Since the data is quite unbalanced (see Figure 4.16) this was used to try to balance the data. This option assigns the weight of each class as inversely proportional to each class' frequencies.

And the parameters chosen for RF were:

- Number of Estimators - sets the number of trees in the forest;
  - Values used = {10, 25, 50, 100, 200, 500}
- Max Depth - limits the length of the longest path from the root to a leaf;
  - Values used = {5, 10, 20}

The tunning of the hyper-parameters was done using the method *Grid Search*. It does an exhaustive search through the previously specified hyper-parameter space. This was achieved in *Python* using *GridSearchCV*[3], a command from the $Scikit-Learn$[4] machine learning package. It creates an estimator class and then passes the hyper-parameters as arguments. This way, all different combinations for the different parameters are tested in a fast and efficient way.

[3] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[4] http://scikit-learn.org/stable/

The best combination of hyper-parameters is then chosen based on various possible *scoring* methods. For this project, $f1\_weighted$ was chosen, which is based on the weighted average of the F1 score of each class. This way the metric accounts for the already mentioned class imbalance, and combines the results across all labels.

For Model Evaluation, *Stratified Cross-Validation* (SCV) (Refaeilzadeh et al., 2009) with 5 folds is used as it is better suited for classification problems (Kohavi et al., 1995). The difference between SCV and standard *Cross-Validation* (CV) relies in the detail that the first tries to ensure that all classes are represented in all folds in approximately the same proportions, while the second splits the data randomly.

SCV was achieved with *GridSearchCV* using the correct parameters. For each fold, the 48 (CART) or the 18 (RF) different hyper-parameters' combinations (8 minimum samples per leaf values $\times$ 6 max depth Figures for CART or 6 number of estimator values $\times$ 3 max depth Figures) are tested (Figure 4.18 step 2) and scored (Figure 4.18 step 3). Next, the score of each hyper-parameter pair is averaged among all the 5 folds, as exemplified in Figure 4.18 step 4.

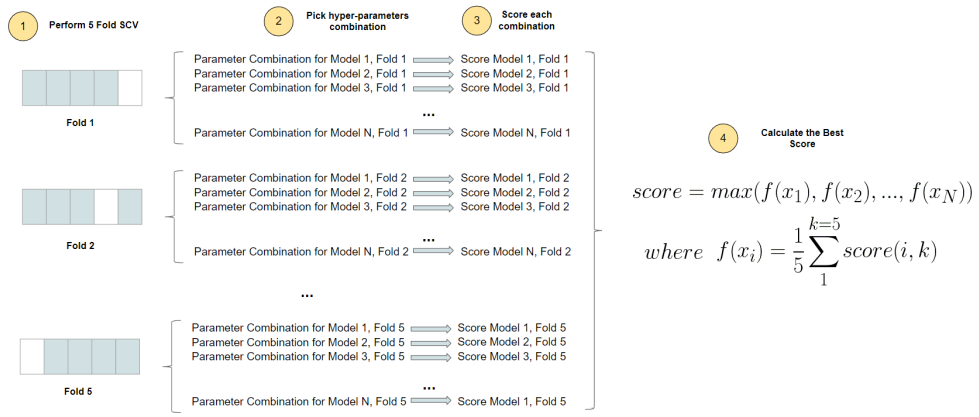


Figure 4.18: GridSearchCV exemplification.

The highest score is then selected, along with its corresponding hyper-parameters. Using those values, a new model is created using the full train data and tested in the test data. This is needed because while doing SCV, the train data is not fully used. Looking at Figure 4.18 step 1, it is evident that for all the train data, only 4/5 - Approximately. Due to the stratified version of CV, the folds might not have the exact same size - of it is actually used as train data, while the other 1/5 - Again, approximately - is used as test data.

4.6.3 Types of Models

The next sections explain the various types of CART models developed in this project. They are described in terms of technique, data split methods and reasoning. Also, some models have two versions where one is created for the whole top 10 cells (global scope), while the second version is created per cell (cell scope), generating in fact 10 different models. A summary of all the developed models is shown in table 4.4.

Type of Model	Scope	Summary
Conventional	Global Cell	Data from all the top 10 cells is used to predict traffic states for all top 10 cells. Data from all the top 10 cells is used to predict traffic states for a single cell at a time.
Sliding Window	Global Cell	Three days from all the top 10 cells are used to predict traffic states for the consecutive day for all top 10 cells. The starting group of three days is then shifted along the month. Three days from all the top 10 cells are used to predict traffic states for a single cell at a time. The starting group of three days is then shifted along the month.
Hierarchical	Global	Data from all the top 10 cells is used to create two models that predict the traffic states for all the cells. The models separate two states at a time, instead of three.
Weekday	Global	Data from each day of the week is used to predict the respective day of the last week.
Combination	Global	A combination of Hierarchical and Weekday.

Table 4.4: Model Summary.

4.6.3.1 Conventional Models

For these models the data at hand was split between train data and test data using as separation point the day 24 of April Any time slot before day 24 is considered training data while anything else beyond that time is test data. This roughly translates into three weeks for training data and one week for test (Figure 4.19).

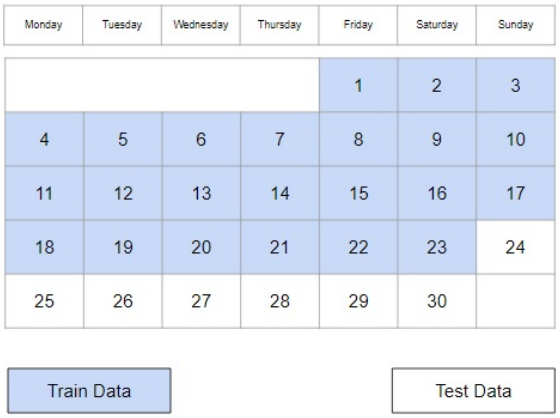


Figure 4.19: Data Split for Conventional Models.

Conventional Global Model - CGM

In this version of conventional model, the full data from the top 10 cells is used to generate a single model that predicts for all cells, the traffic state for each time slot for the last week of April. This allows for an overall understanding of the traffic flow in the top 10 cells.

Due to missing information in some time slots, the number of slots available is not the expected 28800 for the full month (2880 time slots per cell $\times$ 10 cells).The expected values for the train data set and test data set are 22080 (4 slots of 15min per hour $\times$ 24 hours per day $\times$ 23 days $\times$ 10 cells) and 6720 (4 slots of 15min per hour $\times$ 24 hours per day $\times$ 7 days $\times$ 10 cells) respectively.

The actual count of times slots available for each data set is displayed in table 4.5 as well as the percentage of data availability. It is evident that missing information only accounts, at maximum, for less than 4% of the expected value of the train data and less than 9% for the test data. These values, for the grid of 250m, are explained due to the smaller sized cells which contain less information when compared to the greater sized cells.

Data Set	**Grid 250*m***	**Grid 500*m***	**Grid 750*m***
Train Data	21279 (96%)	21819 (99%)	21611 (98%)
Test Data	6175 (92%)	6391 (95%)	6402 (95%)

Table 4.5: Number of Time Slots Available per Grid Size - CGM.

Conventional Cell Model - CCM

For this version of the conventional model, a single model per cell of the top 10 cells is created which predicts the last week of April for each cell. This allows for a better characterization and understanding of traffic flow per cell.

Again, due to missing information, the number of time slots is not the expected 2880 per cell (full month), nor the 2208 and 672 for train and test data sets, respectively. The number of time slots available for each cell is displayed in tables 4.6, 4.7 and 4.8. Again, grid 250*m* accounts for the highest percentage of missing data, reaching 11% on the train data and 17% on the test data for cell 1,30 (hospital area, see Figure 4.8 (a)) with some other cells also missing more than 2% and 4%, train and test data sets, respectively. Also worth noting is cell 5,6 from grid 750*m* reaching almost 15% missing information for train data and 17% on the test data set. This might happen because this is the only cell of this grid without a taxi stand (see table 4.3), which reduces considerably the amount of information generated inside it.

Data Set	**Cell 17,26**	**Cell 17,25**	**Cell 16,26**	**Cell 19,25**	**Cell 17,27**	**Cell 18,26**	**Cell 18,28**	**Cell 16,35**	**Cell 16,27**	**Cell 1,30**
Train Data	2191 (99%)	2175 (99%)	2195 (99%)	2126 (96%)	2168 (98%)	2156 (98%)	2112 (96%)	2004 (91%)	2189 (99%)	1963 (89%)
Test Data	643 (96%)	622 (93%)	640 (95%)	620 (92%)	626 (93%)	623 (93%)	630 (94%)	575 (86%)	641 (95%)	555 (83%)

Table 4.6: Number of Time Slots Available for Grid Size 250m - CCM.

Data Set	**Cell 8,13**	**Cell 8,12**	**Cell 8,14**	**Cell 6,10**	**Cell 7,13**	**Cell 8,11**	**Cell 9,12**	**Cell 9,13**	**Cell 8,15**	**Cell 7,10**
Train Data	2204 (99%)	2197 (99%)	2191 (98%)	2172 (99%)	2196 (99%)	2176 (99%)	2191 (99%)	2182 (99%)	2166 (98%)	2144 (97%)
Test Data	650 (97%)	635 (94%)	648 (96%)	640 (95%)	637 (95%)	624 (93%)	644 (96%)	643 (96%)	641 (95%)	629 (94%)

Table 4.7: Number of Time Slots Available for Grid Size 500m - CCM.

Data Set	**Cell 5,8**	**Cell 5,9**	**Cell 6,8**	**Cell 5,7**	**Cell 5,6**	**Cell 4,9**	**Cell 4,7**	**Cell 6,9**	**Cell 4,6**	**Cell 3,9**
Train Data	2200 (99%)	2204 (99%)	2200 (99%)	2189 (99%)	1885 (85%)	2194 (99%)	2188 (99%)	2170 (98%)	2182 (99%)	2199 (99%)
Test Data	653 (97%)	655 (97%)	649 (97%)	643 (97%)	558 (83%)	647 (96%)	648 (96%)	654 (97%)	647 (96%)	648 (96%)

Table 4.8: Number of Time Slots Available for Grid Size 750m - CCM.

4.6.3.2 Sliding Window

In this model a sliding window technique is used where a small group of days is used as train data to predict the consecutive day, test data. The number of days used in the train data can be selected at will and it is called *Window Length*. For this project, the chosen value of *Window Length* is 3, ensuring at least 27 models, one for each of the 27 predicted days (from the 4^{th} to the 30^{th} of April), as the window is slided across the month of April. Figure 4.20 displays the first two and the last model.

Like in the previous model, there are two scopes for this model: a global version and a cell based version. Both of them allows the understanding of traffic flow per day, while the first focus on the urban area as a whole and the second focus on a cell by cell analysis.

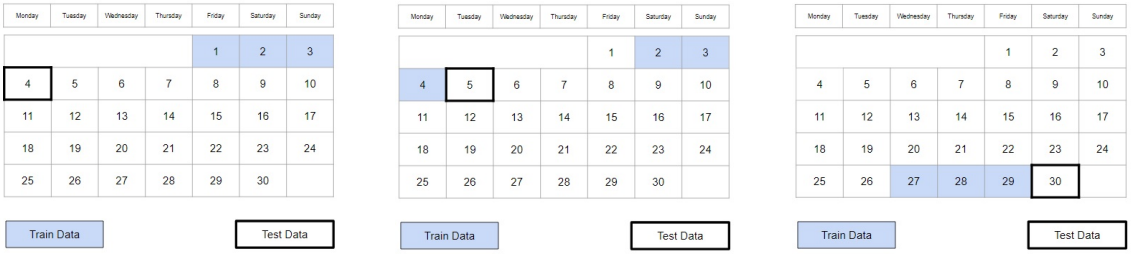


Figure 4.20: Data Split for Sliding Window Model. Figure (a) represents the first model, (b) the second, and (c) is the last built model of the 27 different models

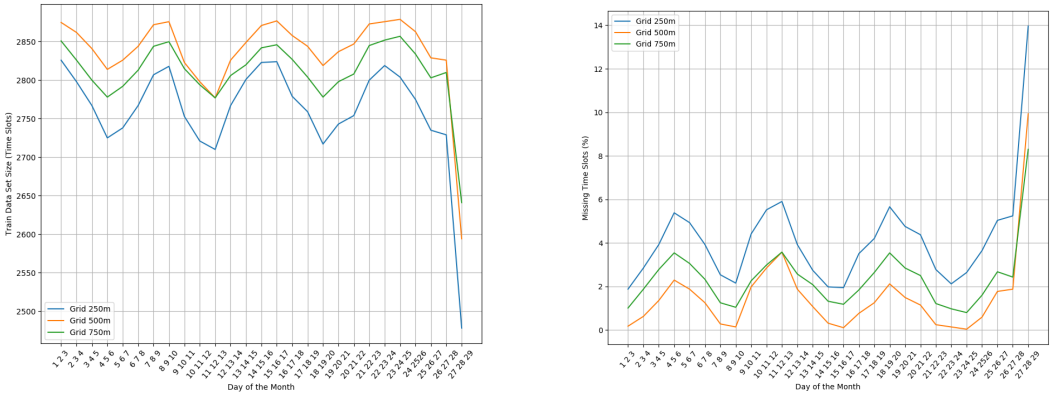
Sliding Window - Global Model

For the global version of this model, the information of the top 10 cells belonging to the sliding window's days is used to predict the next consecutive day for all the cells. In this case, there are 27 different models, one for each day predicted. For each model it is displayed in Figures 4.21 (a) and 4.22 (a) the number of predicted time slots. Again, the value for each data set is lower than expected (2880 for train data - 4 slots of 15min per hour $\times$ 24 hours $\times$ 3 days $\times$ 10 cells and 960 for test data - 4 slots of 15min per hour $\times$ 24 hours $\times$ 10 cells) due to missing information. In these Figures it is also evident the taxi strike (mentioned in section 3.3.1) that affected the whole city, as the quantity of information decreases sharply in both data sets.

Additionally displayed is the overall almost sinusoidal pattern present for all grid sizes. The highest values correspond to groups of day with weekend days, and the smallest values to groups with only week days. The reason for this is related to the city's night activities, where taxis are called and used in late night hours, providing information for time slots that usually have no activity during the week. Although there are more slots with information, it does not mean there is more information at weekends overall. (See section 3.3.1.)

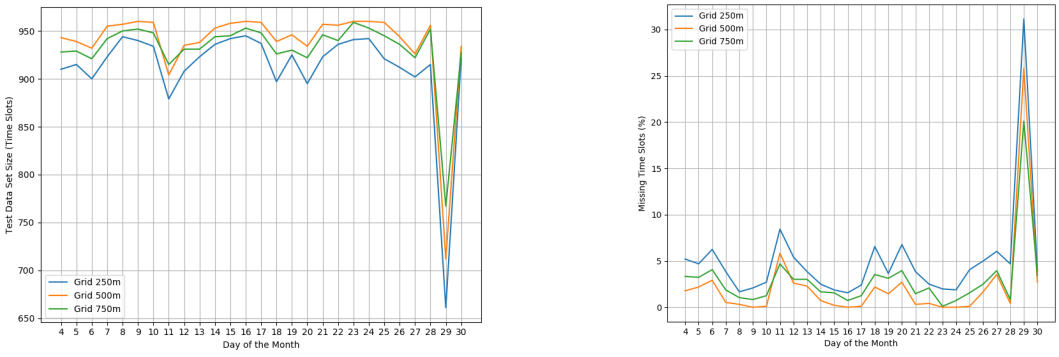
In Figures 4.21 (b) and 4.22 (b) it is displayed the percentage of missing information. Taxi strike also plays a part in the Figures, as the amount of missing info sharply increases on the 29^{th} for the test data or in the train data for the only group of days with the 29^{th} present. The overall sinusoidal pattern is also present, for the same reasons mentioned before.

Excluding the 29^{th}, where missing time slots can go as high as 31% for the grid 250m, the overall percentage does not surpass the 4% for grid 750m, 6% for grid 500m and 8% for grid 250m. This increase in missing information as the grid size decreases is expected as the information per cell decreases per cell.



(a) (b)

Figure 4.21: Number of Predicted Time Slots for the Train Data and respective missing data percentage - Sliding Window Global Model.

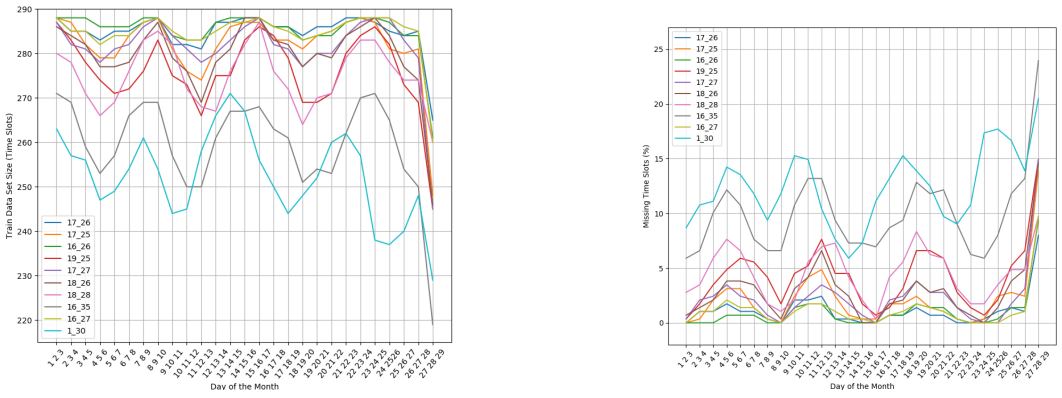


(a) (b)

Figure 4.22: Number of Predicted Time Slots for the Test Data and respective missing data percentage - Sliding Window Global Model.

Sliding Window - Cell Model

In this scope of the sliding window, it is created a single model per cell of the top 10 cell. A group of three days is used as train data to predict the consecutive day, considered now test data. In total there were created 270 different models, 27 per cell for each day. The expected number of time slots is 288 for the train data (4 slots of 15*min* per hour $\times$ 24 hours $\times$ 3 days) and 96 for the test data (4 slots of 15*min* per hour $\times$ 24 hours). Actual values, due to missing information, are shown in Figures 4.23 and 4.24 for the example of grid 250*m*. Aside from cells 1,30 and 16,35, the missing information percentage does not go above 7% for the train data in Figure 4.23 (b), with the exception being the prediction for day 30, because, as said before, it uses the 29^{th} which has a very low quantity of data. Cells 1,30 and 16,35 present a fairly high percentage missing of train data, because there are the two cells outside of the urban area of the city. Even though they are located at a hospital (cell 1,30) and in a main train station (cell 16,35), these points of interest represent a small amount of activity when compared to the center of the city. As for the test data values shown in Figure 4.24, the same scenario is displayed, with cells 16,35 and 1,30 missing up to 18 and 25% respectively, while the other cells barely going above 10%.



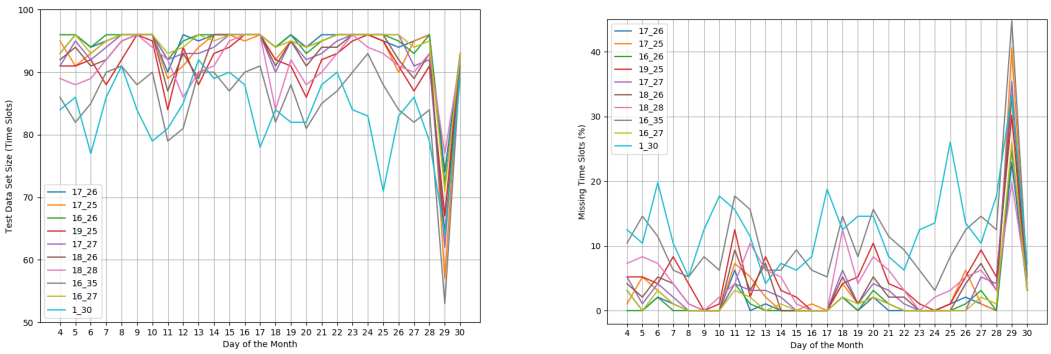
(a) (b)

Figure 4.23: Number of Predicted Time Slots for the Train Data and respective missing data percentage per cell for Grid Size 250m - Sliding Window Cell Model.

4.6.3.3 Weekday-Based

The Weekday-based model creates 7 different models, one for each day of the last week of April using as train data, the same days of the previous weeks are used. Looking at Figure 4.25 it is possible to see the data split between train and test data for each day of the week. The April 25 and 29 are not used, because as mentioned before, the first is a holiday and the second had a taxi strike, therefore these days have a very different behavior when compared to the others (see Figure 3.5).

The expected number of time slots is different for each day of the week, because each day takes in a different number of training days. Monday takes the least with only 2 days of train data



(a) (b)

Figure 4.24: Number of Predicted Time Slots for the Test Data and respective missing data percentage per cell for Grid Size 250m - Sliding Window Cell Model.

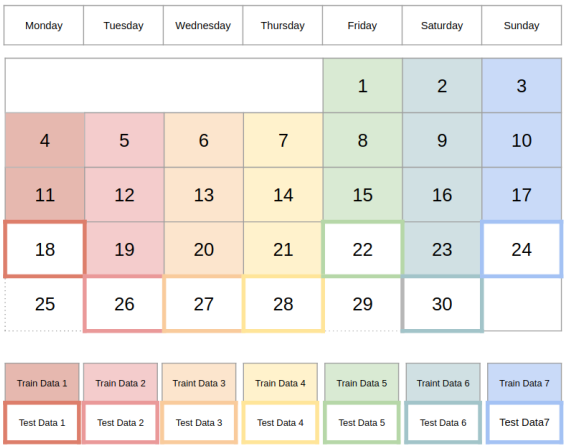
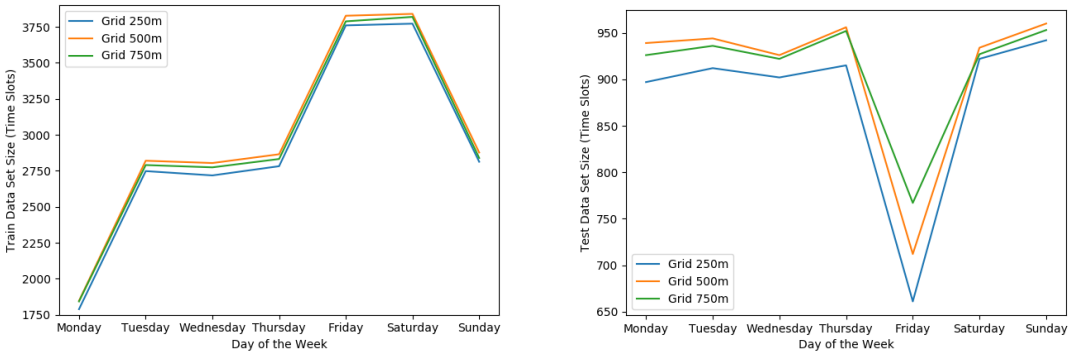


Figure 4.25: Data split for the Weekday Model.

(1920 expected time slots - 4 slots of 15*min* per hour × 24 hours × 2 days × 10 cells). Tuesday, Wednesday, Thursday and Sunday take 3 days of train data (2880 expected time slots - 4 slots of 15*min* per hour × 24 hours × 3 days × 10 cells). And lastly, Friday and Saturday have each 4 days of train data (3840 expected time slots - 4 slots of 15*min* per hour × 24 hours × 4 days × 10 cells). All different models are expected to predict a single day (960 expected time slots - 4 slots of 15*min* per hour × 24 hours × 10 cells). Looking at Figures 4.26 and 4.27 it is possible to see the actual number of predicted slots and missing information percentage, respectively.

In Figure 4.26 (a), the number of slots is overall the same for the grid sizes, with no distinguishable difference. In Figure (b), though, it is easy to notice how grid 250*m* has a lower amount of time slots due to the smaller sized cells. Friday also presents a great break in the amount of available time slots, due to the taxi strike mentioned in section 3.3.1, on the April 29, 2016.

Looking at Figure 4.27 (a), the overall percentage of missing time slots is below 4% with grid 250*m* again presenting higher values, specially at Monday with 7%. The overall decrease in missing time slots is due to night activities during the weekend, where more people use taxis at late night hours, that typically have almost no taxi trips during the week. This way, the number of time slots with information increases during the weekend. Lastly in Figure 4.27 (b), aside from the expected higher overall value of grid 250*m* missing information, Friday has a peak, due to the already mentioned taxi strike.

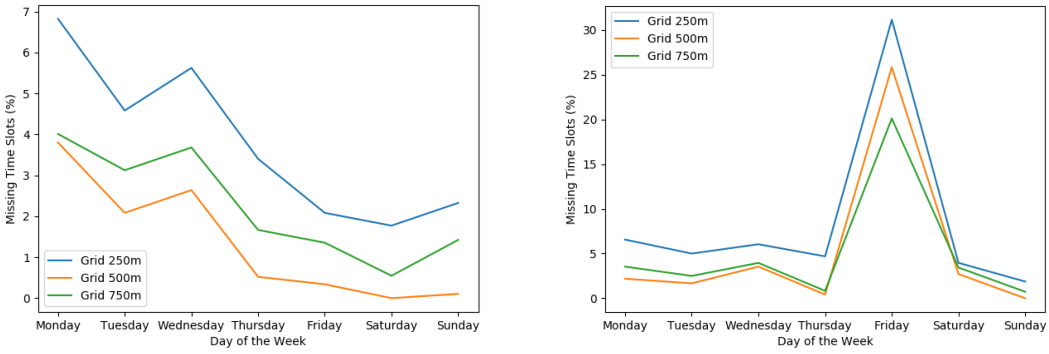


(a) (b)

Figure 4.26: Train and Test data sets' size for the different day of the week.

4.6.3.4 Hierarchical

In this model, instead of creating a single model that predicts all traffic states, two models are created to predict the separation between classes, similar to Wang and Casasent (2009) and reviewed in detail in Silla Jr and Freitas (2011). Looking at Figure 4.28, two models are trained. The first (Model 1) to separate between a pre-specified class (Class 1), called Macro Class, and any other class (Non Class 1), and the second (Model 2) to separate the classes of Non Class 1 in Class 2 and Class 3.



(a) (b)

Figure 4.27: Train and Test data sets' missing slots percentage for the different days of the week.

After training the models, they are applied to the data at hand. The model 1 is applied to the full test data set predicting the pre-specified class and non class. The model 2 is only enforced on the Non Class group generated before, predicting the two other classes inside it. Since this project only dealt with three different classes, all combinations of macro classes and respective separations were attempted.

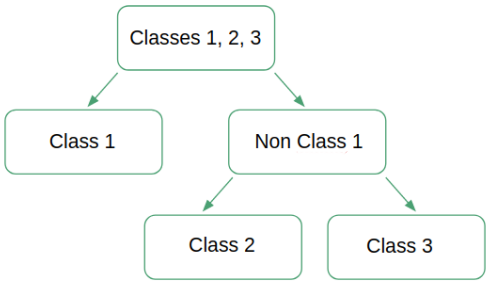


Figure 4.28: Diagram of Hierarchical Model.

4.6.3.5 Model's Combination

In this model a combination of two previous models was used: Hierarchical and Weekday-based. First, each weekday is grouped and the aim is to predict the corresponding day of the last week, just like in weekday-based model. Then, for each day it is created 2 models instead of only one. The first model separates the selected macro class from the other two, and the second models separates the last remaining classes, just like in the hierarchical model.

The train and test data split is the same as in Figure 4.25. What changes is the number of time slots to predict, as there are actually two groups to prediction, one for each split in classes, according to Figure 4.28.

4.7 Feature Extraction

Feature extraction is one of the last steps of this project, where each feature's contribution to congestion or other traffic states is evaluated. The weight of each feature for each model is measured as well as their interpreted in the generated decision tree, in order to better understand the mechanics of traffic congestion and its causes.

4.8 Evaluation Metrics

In order to compare between the different types of models, various types of metrics are used depending on their purpose. To evaluate the classification, three types of F1 scores were used: Micro Average, Macro Average and Weighted Average. This will give a base for comparison in terms of predictability. As for the features chosen by each model as their most important, the metric chosen is Gini Importance.

The F1 score is a measure of a test's accuracy which uses two parameters: Precision and Recall. The difference between the three mentioned types of F1 score is the way these constants are calculated. The overall formula for the F1 score is seen in equation (4.1). Gini Impurity measures the impurity of each leaf of the decision tree, with the purpose of decreasing it at every split.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.1}$$

4.8.1 Micro Average - *mF1*

The Micro F1 score calculates precision and recall by summing up all the true positives (TP), false positives (FP) and false negatives (FN) across all classes. It gives the same weight to all classifications, so a class with a larger F1 score will influence the metric more than the F1 scores of a smaller class. The formulas used are shown in equations (4.2) and (4.3) for precision and recall, respectively. Index i represents each class while k is the number of total classes, 3.

$$Precision = \frac{\sum_i^k TP_{\mathrm{i}}}{\sum_i^k TP_{\mathrm{i}} + \sum_i^k FP_{\mathrm{i}}} \tag{4.2}$$

$$Recall = \frac{\sum_i^k TP_{\mathrm{i}}}{\sum_i^k TP_{\mathrm{i}} + \sum_i^k FN_{\mathrm{i}}} \tag{4.3}$$

4.8.2 Macro Average - *MF1*

The Macro F1 score calculates precision and recall per class using the formulas (4.4) and (4.5), respectively, attributing equal weights to each class. Index i represents each class while k is the number of total classes, 3.

$$Precision = \frac{1}{k}\sum_{i}^{k}\frac{TP_{\mathrm{i}}}{TP_{\mathrm{i}}+FP_{\mathrm{i}}} \tag{4.4}$$

$$Recall = \frac{1}{k}\sum_{i}^{k}\frac{TP_{\mathrm{i}}}{TP_{\mathrm{i}}+FN_{\mathrm{i}}} \tag{4.5}$$

4.8.3 Weighted Average - *aF1*

The Weighted Average F1 score takes into account the total number of instances, unlike the previous metrics. The formulas for weighted precision and recall are 4.6 and 4.7, respectively. I and TI represent the number of total instances and the number of true instances of a class, respectively. This ensures a weighted factor depending on the size of the class, when calculating the precision and recall.

$$Precision = \frac{1}{\#I}\sum_{i}^{k}\#TI_{\mathrm{i}}P_{\mathrm{i}} \tag{4.6}$$

$$Recall = \frac{1}{\#I}\sum_{i}^{k}\#TI_{\mathrm{i}}R_{\mathrm{i}} \tag{4.7}$$

4.8.4 Gini Importance

Gini Importance, or impurity, is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. When used in a decision tree, the algorithm experiments splitting the data in all possible combinations for each feature in order to find the lowest possible value of gini impurity (highest value of purity), reducing the cost of misclassification, with the purpose of reaching a single class per leaf (Breiman, 1996). The formula to calculate it is:

$$Gini = 1 - \sum_{j=1}^{c} {p_{\mathrm{j}}}^2 \tag{4.8}$$

Where p_j is the probability of an item with class j being chosen, c represents the number of classes and $j \in \{1, 2, 3\}$. Feature weight is measured as the normalized total reduction of the gini importance.

Chapter 5

Results and Discussion

5.1 Chapter overview

In this chapter the results for each model are displayed, both in terms of predicting capabilities and interpretation. Results for the different metrics are shown as well as the selected features per model and respective decision trees.

5.2 Results

5.2.1 Conventional Model

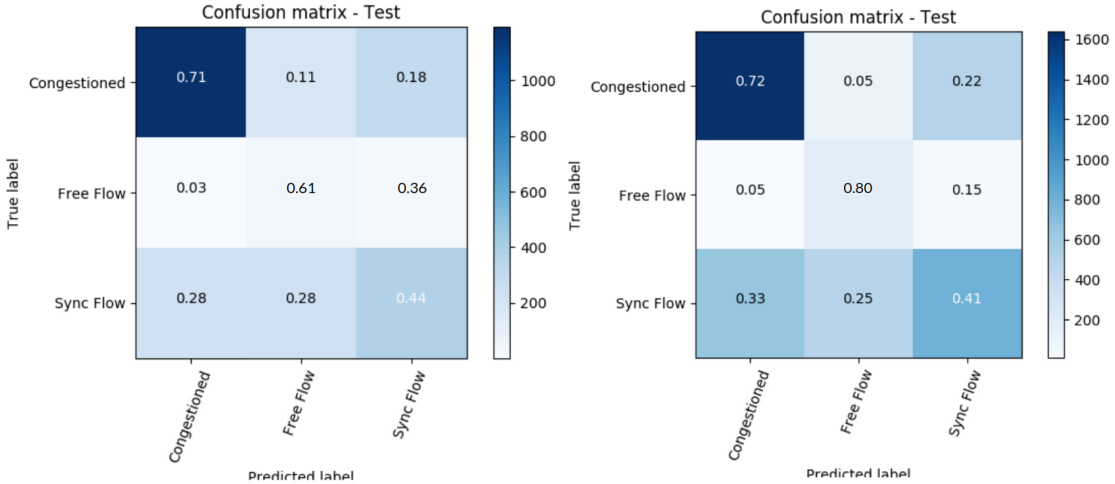
Table 5.1 presents the *micro F1 (mF1)*, *Average F1 (aF1)* and *Macro F1 (MF1)* scores for the Conventional Global Model (CGM). Column *Score* refers to the result that the best model received when performing hyper-parameter tunning on the training data, along with its standard deviation. Column *Test Results*, displays the metrics when the best model is applied to the test data and *Running Time* is the processing time for that grid size. As expected the results tend to be better for higher grid size models, due to greater amount of data overall.

	Trainer Results				Test Results			
Grid Size (m)	Score $\pm$ SD	*mF1*	*aF1*	*MF1*	*mF1*	*aF1*	*MF1*	Running Time (s)
250	0.55 $\pm$ 0.06	0.66	0.63	0.51	0.61	0.58	0.47	366
500	0.51 $\pm$ 0.16	0.62	0.62	0.52	0.59	0.58	0.52	577
750	0.55 $\pm$ 0.05	0.68	0.65	0.60	0.61	0.58	0.54	508

Table 5.1: Conventional Model Results - Global Scope

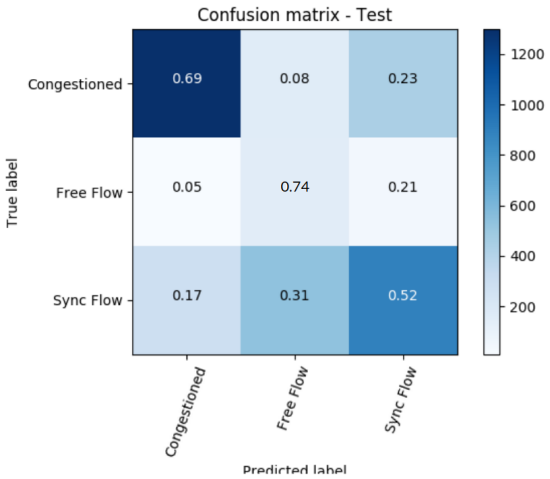
Also displayed are the normalized confusion matrices for each model in Figure 5.1, displaying the performance of the model for each grid size on the test data. It's apparent on 5.1 (a) that the *Sync Flow* state is frequently being confused with the other states (5.1 (b) and (c)) but also that *Congestion* is very easily detected, since it is correctly classified almost 70% across all grid sizes. On Figure 5.1 (b) the *Sync Flow* is heavily confused again with the other states, with *Congestion* starting now to be confused with *Sync Flow*. Lastly, in 5.1 (c), *Sync Flow* state is more easily

separated from the other states, while *Free Flow* remains to be greatly confused with the other states. The worst performance of the smaller grid size is once again happening due to the smaller amount of information per cell, per time slot, providing less data to train and test the model. In Figure 5.1 (c), the states achieve better results due to greater amount of information, which allows smaller classes like *Free Flow* to gain representation and further balance the data.



(a) Grid 250m

(b) Grid 500m



(c) Grid 750m

Figure 5.1: CGM - Normalized Confusion Matrices.

After this overview of the global scope model and in order to better understand what is happening on each grid size, each cell had its own unique model created. The results are displayed in tables 5.2, 5.3 and 5.4, along with the comparison with the global scope of the model. In green are the metrics that improved when compared to the previous version, showing as expected that the majority of the cells benefited from creating its own model as opposed to a more general one. This happens because the majority of the cells have its own traffic model which is sometimes different than the global network model.

Metrics	Global	Cell 17,26	Cell 17,25	Cell 16,26	Cell 19,25	Cell 17,27	Cell 18,26	Cell 18,28	Cell 16,35	Cell 16,27	Cell 1,30
mF1	0.61	0.87	0.60	0.57	0.80	0.67	0.54	0.51	0.71	0.62	0.50
aF1	0.58	0.86	0.59	0.55	0.80	0.66	0.52	0.48	0.67	0.60	0.47
MF1	0.47	0.54	0.36	0.35	0.56	0.44	0.52	0.40	0.45	0.54	0.38

Table 5.2: Conventional Model Results, Grid 250m - Global and Cell Scopes.

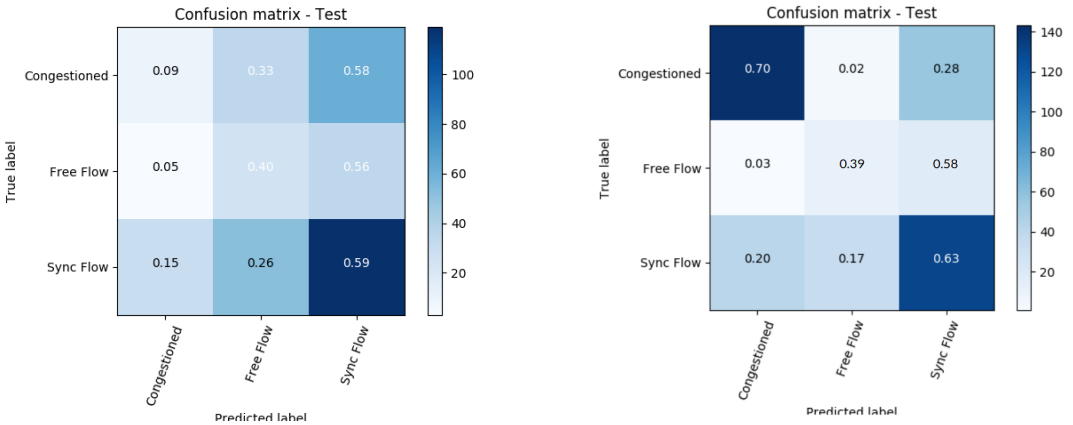
Metrics	Global	Cell 8,13	Cell 8,12	Cell 8,14	Cell 6,10	Cell 7,13	Cell 8,11	Cell 9,12	Cell 9,13	Cell 8,15	Cell 7,10
mF1	0.59	0.82	0.63	0.65	0.69	0.56	0.65	0.73	0.69	0.61	0.53
aF1	0.58	0.80	0.62	0.64	0.68	0.54	0.64	0.73	0.68	0.60	0.52
MF1	0.52	0.71	0.48	0.43	0.49	0.47	0.55	0.69	0.55	0.52	0.49

Table 5.3: Conventional Model Results, Grid 500m - Global and Cell Scopes.

Metrics	Global	Cell 5,8	Cell 5,9	Cell 6,8	Cell 5,7	Cell 5,6	Cell 4,9	Cell 4,7	Cell 6,9	Cell 4,6	Cell 3,9
mF1	0.61	0.74	0.79	0.75	0.65	0.41	0.60	0.65	0.62	0.52	0.59
aF1	0.58	0.71	0.79	0.73	0.64	0.43	0.58	0.62	0.61	0.51	0.58
MF1	0.54	0.62	0.52	0.59	0.58	0.33	0.39	0.51	0.58	0.44	0.48

Table 5.4: Conventional Model Results, Grid 750m - Global and Cell Scopes.

The cells in which the metrics decreased, are cells that have a small amount of data (e.g.: Cells 1,30, 7,10 and 3,9 for grid sizes 250*m*, 500*m* and 750*m*, respectively) or have a very small amount of at least one traffic state, causing severe unbalance on the data that cannot be corrected with the adopted methods (See section 4.6.2), lowering the value for the Macro F1 metric (e.g.: Cell 16,26, 7,13 and 5,6 for grid sizes 250*m*, 500*m* and 750*m*, respectively). Figure 5.2 (a) shows, as an example, the effect that a low amount of data (Cell 5,6 is the only one in the grid size 750*m* without any taxi stand - see table 4.3 in section 4.3.1.1.) can have in the cell's model. More than 70% of all the *Congestion* state and 65% of the *Free Flow* are being classified as *Sync Flow*. On the other hand, looking at 5.2 (b), which represents cell 8,13 of grid size 500*m* (cell with the highest amount of data for this grid size), now 70% of the *Congestion* state is well classified and the misclassification of *Free Flow* state has also lowered.

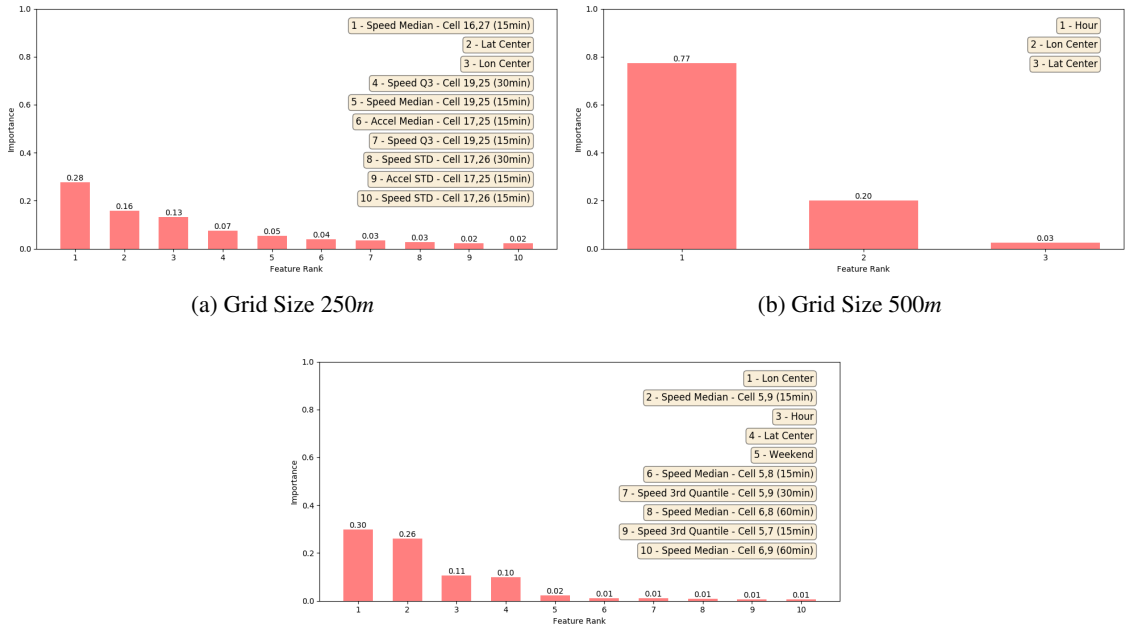


(a) Confusion Matrix - Cell 5,6 - Grid Size 750*m*

(b) Confusion Matrix - Cell 8,11 - Grid Size 500*m*

Figure 5.2: Comparison between Confusion Matrices - Conventional Model.

In order to better separate between different states, the hierarchical method was implemented, and will be evaluated in the next section.



(a) Grid Size 250m

(b) Grid Size 500m

(c) Grid Size 750m

Figure 5.3: Top Features' Importance for all Grid Sizes - Conventional Global Model.

Focusing now on the selected features and their importance, Figure 5.3 (a) represents the feature importance for the top 10 (more than 1% weight) features for grid size 250m, and it shows the most import features are the speed and acceleration statistics such as median, 3^{rd} quartile and standard deviation of the cells with the most information (17,26 and 17,25) for the last 15 *min*. Also worth noting are the coordinates selected, displaying a congestion state in the center of the urban area (all top 10 but cells 1,30 and 16,35) in the decision tree.

On grid size 750m (Figure 5.3 (c)) the longitude and latitude gain more importance than in grid 250m, showing again an increased amount of congestion state in the center of the urban area. The feature *hour* appears with some importance (0.11) giving the information from the decision tree that before 7h in the morning there's an increased amount of *Free Flow* state through all top 10 cells and also that 18h is the hour at which the congestion typically decreases for the urban area (cells 5,6, 5,7, 5,8, 5,9, 6,8 and 6,9). Also worth mentioning is the increase in interval under which previous speed and acceleration statistics of cells still hold importance (from 15$\sim$30 *min* in grid size 250m to 15$\sim$60 *min* for this grid size). Since the cells are larger, it is normal that whatever traffic conditions that flows through them, takes longer to propagate, unlike grid size 250m.

Lastly, on grid size 500m, only three features are selected: hour, latitude and longitude (Figure 5.3). The decision tree displayed in 5.4 shows that for hours before 6 in the morning, the traffic is mainly in the *free flow* state for all cells. Also, 18h in the afternoon is the point after which the congestion tends to reduce for the main center area of the urban area (cells 7,13, 8,12, 8,13, 8,14,

9,12 and 9,13).

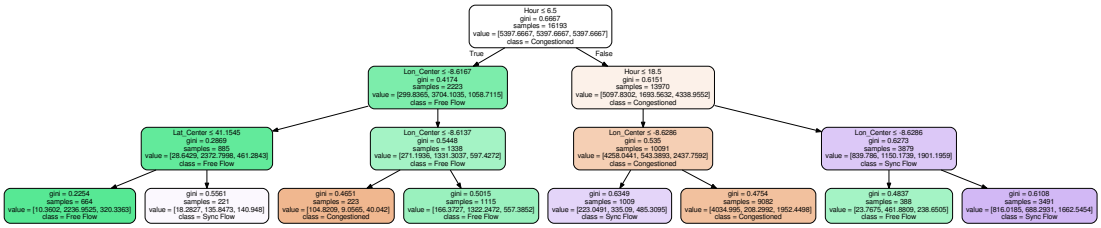
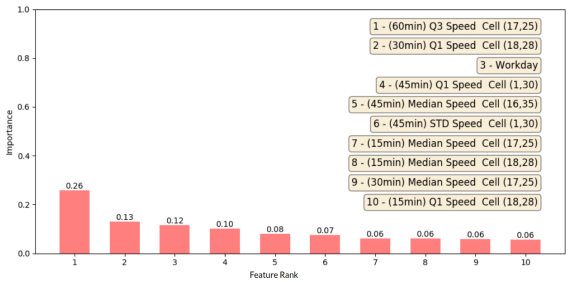
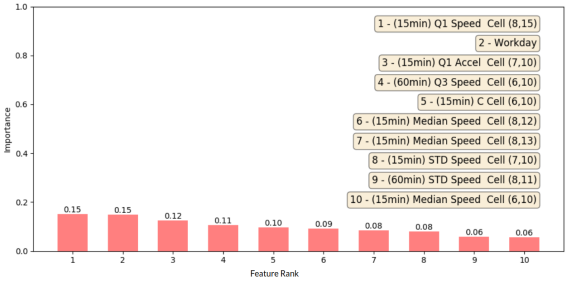


Figure 5.4: Representation of the Decision Tree for Conventional Global Model, Grid Size 500m.

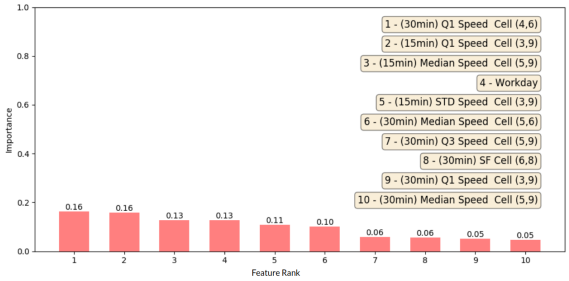
As for the cell based models, the top 10 features' importance for all grid sizes are displayed in Figure 5.5. Overall, the importance of the *weekday* feature increases when compared to the global scope, reaching weight 0.12, 0.15 and 0.13 for grid sizes 250*m*, 500*m* and 750*m* respectively. For grid size 250*m*, the most important features become the last hour and half hour for cells 17,25 and 18,28, respectively, which are located outside the center urban area, explaining why longer intervals are more important to estimate the state of the center, as the traffic takes longer to reach it. In grid size 500*m* the speed and acceleration features (1^{st} quartile for both, and 3^{rd} quartile for speed) gain further importance for cells 8,15, 6,10 and 7,10, which have in them main roads that lead to the urban center. Also, the median and standard deviation for the two cells with most information (cells 8,12 and 8,13) become important, most likely due to their greater amount of information. For grid size 750*m* cells that are far away from the center (cells 4,6 and 3,9) become important again, as well as the one the cells with most information (cell 5,9).



(a) Grid Size 250*m*



(b) Grid Size 500*m*



(c) Grid Size 750*m*

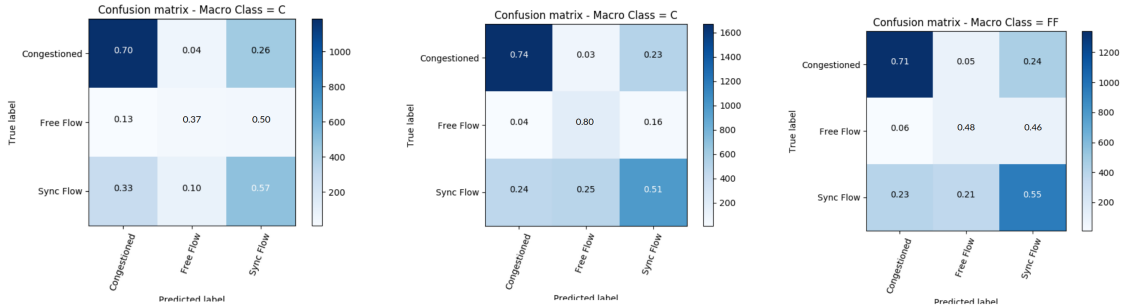
Figure 5.5: Top Features' Importance for all Grid Sizes - Conventional Cell Based Model.

5.2.2 Hierarchical Model

An attempt to better separate between classes is done on this model, by creating two specific models that are trained to separate between only two classes, instead of three at the same time. Table 5.5 shows the results for the performance of this model on the test data and it displays an overall increase on the Macro F1 metric, which means the classes are not being misclassified as much as they used to in the previous model. Further confirmation of this can be seen in Figures 5.6 where not only all grid sizes maintained a correct classification of both *Congestion* and *Sync Flow* above 70% and 50% respectively, but also the *Free Flow* has reached a maximum of 80% correctly classified for the grid size 500*m*, when using *Congestion* as the macro class.

	Grid Size 250*m*				Grid Size 500*m*				Grid Size 750*m*			
Macro Class	*mF1*	*aF1*	*MF1*	Running Time (*s*)	*mF1*	*aF1*	*MF1*	Running Time (*s*)	*mF1*	*aF1*	*MF1*	Running Time (*s*)
Congestion	0.65	0.66	0.50	478	0.64	0.65	0.56	886	0.66	0.67	0.55	805
Free Flow	0.66	0.67	0.45	553	0.64	0.65	0.50	933	0.71	0.70	0.54	793
Sync. Flow	0.63	0.65	0.48	598	0.60	0.62	0.53	1075	0.63	0.64	0.53	940

Table 5.5: Hierarchical Model's Results.



(a) Confusion Matrix - macro class *Congestion*, Grid Size 250*m*.

(b) Confusion Matrix - macro class *Congestion* - Grid Size 500*m*.

(c) Confusion Matrix - macro class *Free Flow* - Grid Size 500*m*.

Figure 5.6: Comparison between Confusion Matrices - Hierarchical Model.

Displayed in Figures 5.7 are the features' importances for the different grid sizes. The feature's importance on this model is average through all models for all macro classes. It is shown that mainly some specific cells contribute to the traffic states and how they are propagated trough the other cells.

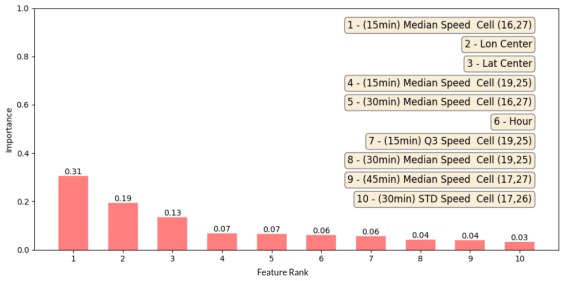
Figure 5.7 (a) shows grid size 250*m* and the cells that mainly contribute for this model are 16,25 and 19,27, with their different speed statistics (Standard Deviation, 1^{st}, 2^{nd} and 3^{rd} quartiles) for the previous 15, 30, 45 and 60 minutes.

Figure 5.7 (b) displays the second model for the macro class *Sync Flow* in grid size 500*m*, where it separates synchronized flow and congestioned states. The cells that provided the most important features are 8,11 and 9,12, with their 1^{st} and 3^{rd} quartiles for the previous 15 minutes.

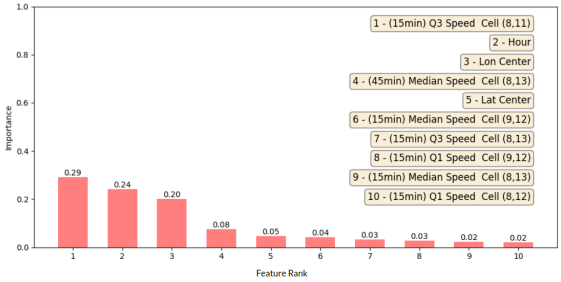
Lastly, Figure 5.7 (c) presents the first model for the macro class *Congestion* for grid size 750*m*, where the congestion state is separated from the other two classes. The cells that help the

most in creating this model are 5,9 and 5,7 with their Median and 3^{rd} quartile for the previous 15 minutes.

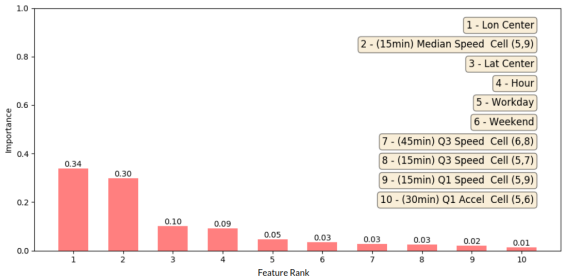
When looking at the location of each cell, it is apparent that they are not located in the center of the urban area, but instead around it, with the main roads that lead to the center, inside them. They are even located in roughly the same areas across the grid sizes: Cells 16,27 from grid size 250m and 5,9 from grid size 750m have some of the main roads that lead to the center of the city, coming from the east. Cells 19,25 from grid size 250m and 9,12 from grid size 500m share some of the roads that take the traffic from near the river to the center area when coming from the southwest direction. Lastly, cells 8,11 from grid size 500m and 5,7 from grid size 750m contain the main road for the traffic coming from the east. These results show that the cells around the urban center are the most important when predicting its traffic state.



(a) Top Feature's Importance, Grid Size 250m



(b) Top Feature's Importance, Grid Size 500m



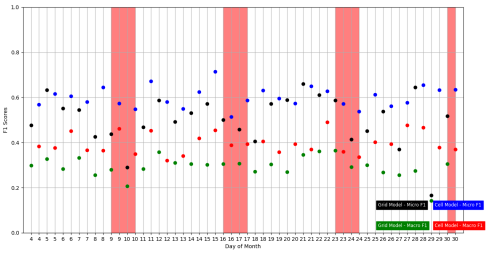
(c) Top Features' Importance, Grid Size 750m

Figure 5.7: Feature's Importance - Hierarchical Model.

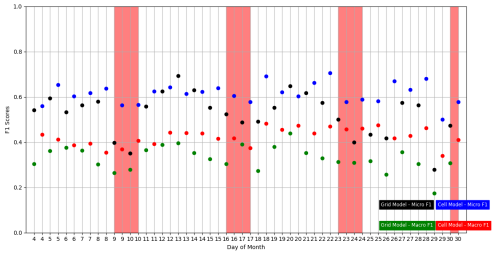
5.2.3 Sliding Window

On this approach, each day was predicted using as train data the previous three days. This resulted in 27 different models for the global scope version, one for each day of April, starting on the 4^{th}. The results for the test data can be seen in Figures 5.8 where it shows for all grid sizes, the comparison between the two scopes, global and cell based, for metrics Micro and Macro F1. For presentation's sake, instead of displaying the all the results for each scope, only the median is displayed. This ensures that the single shown value is representative of the full list of results.

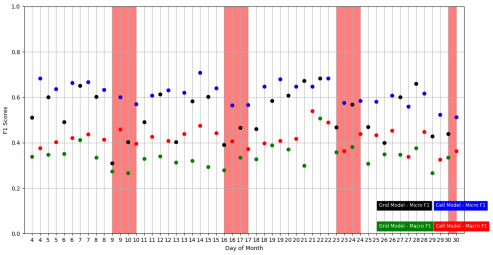
Analyzing all Figures in 5.8 shows the overall idea that cell based models provide better results across all metrics with only a few exceptions (e.g.: 5^{th} of April, Micro F1 score for grid size 250m,



(a) Grid 250m



(b) Grid 500m



(c) Grid 750m

Figure 5.8: Sliding Window Median Results for all three grid sizes. Red sections represent weekends.

13^{th} of April, Micro F1 score for grid size 500m and 27^{th} of April, Macro F1 score for grid size 750m). The overall improve with the change of scope is due to certain cells having their own traffic model which is difficult to predict with a full grid based model. The exceptions to this can be explained by exceptional events on that cell, like an accident or construction work. If these have a considerable duration during a day, they might affect the results for that cell. Also shown is an overall sinusoidal pattern for all metrics and all scopes. This happens due to an overall decrease of metrics over the weekend days (in red) caused by using the previous three weekdays to predict a weekend day (example, using the 6^{th} (Wednesday), 7^{th} (Thursday) and 8^{th} (Friday) to predict the 9^{th} (Saturday)). As shown before in section 3.3.1, there are quite a few differences between the two groups of days, and so the model trained on those three days will retrieve lower values across all metrics when predicting a weekend day.

	Grid Size 250m				Grid Size 500m				Grid Size 750m			
Scope	*mF1*	*aF1*	*MF1*	Running Time (s)	*mF1*	*aF1*	*MF1*	Running Time (s)	*mF1*	*aF1*	*MF1*	Running Time (s)
Grid Based	0.50	0.50	0.29	1225	0.52	0.52	0.33	1869	0.52	0.52	0.34	1677
Cell Based	0.60	0.59	0.40	3666	0.62	0.60	0.42	4398	0.62	0.64	0.42	4121

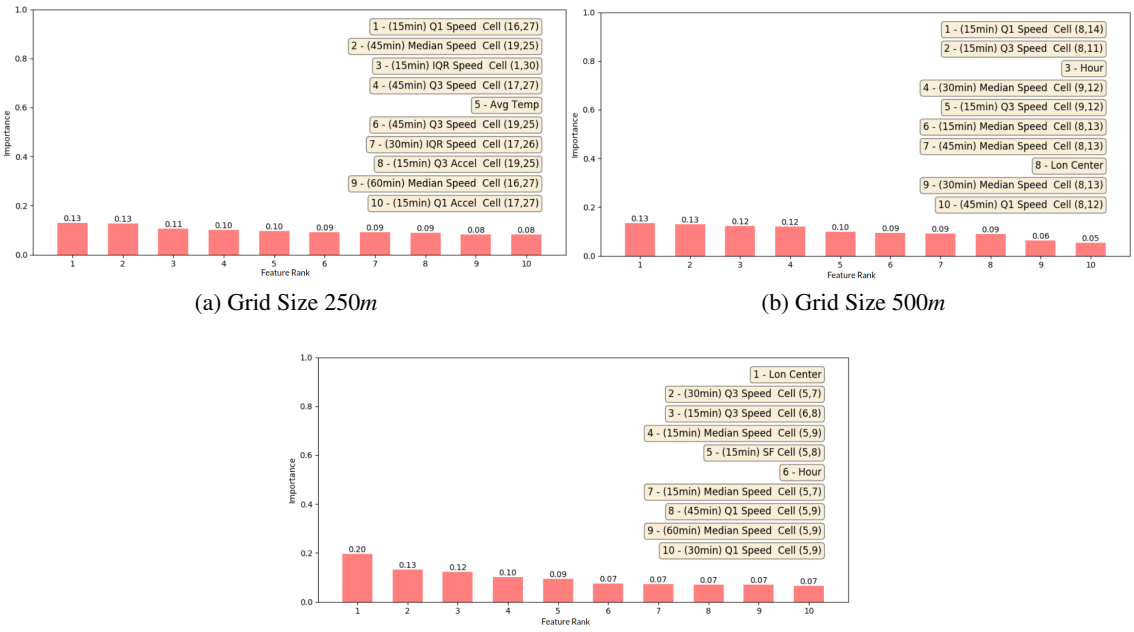
Table 5.6: Sliding Window Results - Mean Median.

Table 5.6 shows the mean values of the results displayed in 5.8. It demonstrates the increase across all metrics of about 10% when using cell based model. But there are traffic patterns that are not repeated throughout the week days, but are instead happening at specific days, like a fair that only happens at Thursdays or a specific congestion happening only at Mondays. These type of weekday patterns will be explored in the next section, weekday models.

As for the feature importance, the average weight of each feature is displayed in Figures 5.9. In Figure 5.9 (a), the features average importance is quite evenly distributed, without a feature standing out. The features selected are the speed and acceleration statistics from various different cells, but with a small incidence on cells 16,27 and 19,25. Also interesting is the interval under which each cell provides information considering great values like 45 and 60*min*, as well as the average temperature having a considerable importance (0.10).

Grid size 500*m* and its features are displayed in Figure 5.9 (b), where again no feature stands out, but rather the focus on urban cells and their speed statics (1^{st}, 2^{nd} *and* 3^{rd} quartiles) such as 8,11, 8,12, 8,13 and 8,14. Also relevant is the interval for the collection of information increasing for the less important features.

Lastly in Figure 5.9 (c), it is shown that longitude holds the highest importance, followed by cells 5,7, 6,8 and specially 5,9, providing different speed statistics (1^{st}, 2^{nd} *and* 3^{rd}) to build the model. Also, the *Synchronized Flow* state of cell 5,8 (the cell with the highest amount of information for this grid size) has 0.09 of importance when creating the model and predicting the traffic state.



(a) Grid Size 250*m*

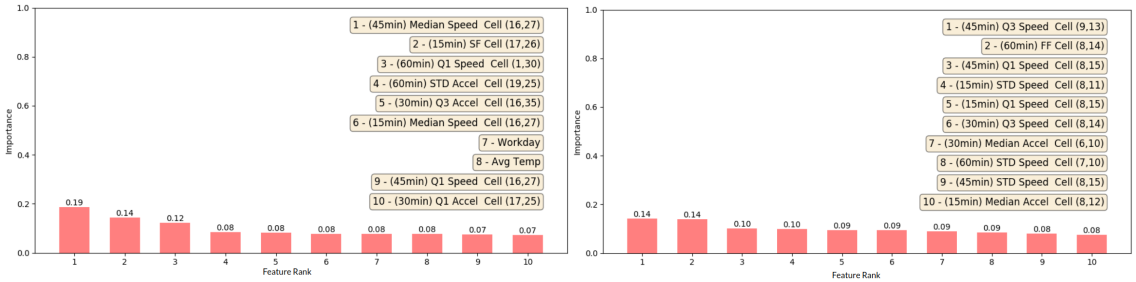
(b) Grid Size 500*m*

(c) Grid Size 750*m*.

Figure 5.9: Top 10 Features' Average Importance - Sliding Window Global Model.

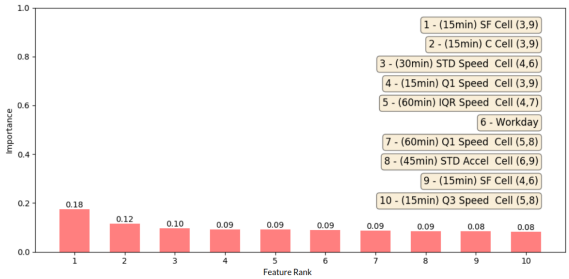
Changing the scope for a cell based model of the sliding window method, the results are displayed in Figures 5.10 and are inconclusive as no feature stands out again for any of the grid sizes. Figure 5.10 (a) represents the top 10 features' weight for grid size 250*m*, and the most relevant features are the *synchronized flow* state of cell 17,26. Other features are speed and acceleration statistics such as 1^{st}, 2^{nd} *and* 3^{rd} quartiles for cells 1,30, 16,27, 19,25 and 17,25. Also the intervals under which cell provides relevant information is quite large, reaching 60 minutes frequently. In Figure 5.10 (b) for grid size 500*m*, the intervals that provide important information are large again,

and focus again on the speed and acceleration statistics of various cells such as 9,13, 8,14, 8,15 and 8,11. Lastly on grid size 750*m*, displayed in Figure 5.10 (c), the two important features are the *synchronized* and *congestion* states of the cell with less information on this grid size, possibly due to having some of the roads that lead to the center of the urban area, when coming from the north.



(a) Grid Size 250*m*

(b) Grid Size 500*m*



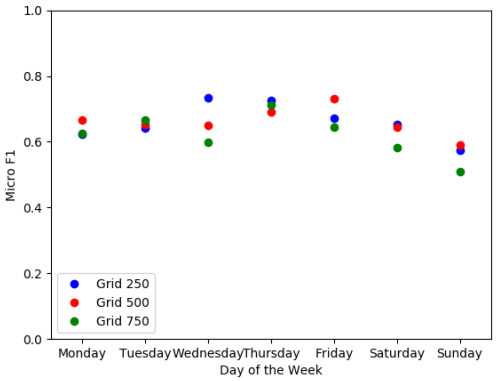
(c) Grid Size 750*m*.

Figure 5.10: Top 10 Features' Average Importance - Sliding Window Cell Based Model.

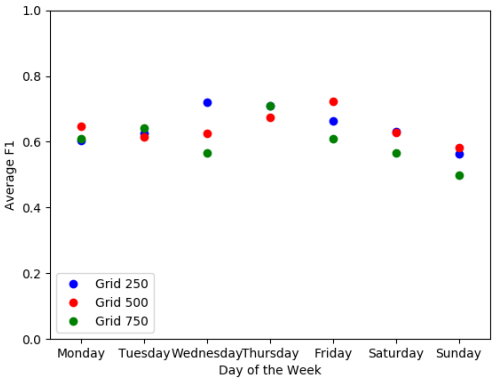
5.2.4 Weekday

On this approach, seven models were trained, one for each day of the week. The results for the test data, obtained for each metric are displayed in Figures 5.11. Figure 5.11 (a) shows the Micro F1 variation throughout the week for the different grid sizes, displaying a lower value for both Saturday and Sunday for Micro and Average F1. In general, Macro F1 tends to have lower values during the week than at weekend.

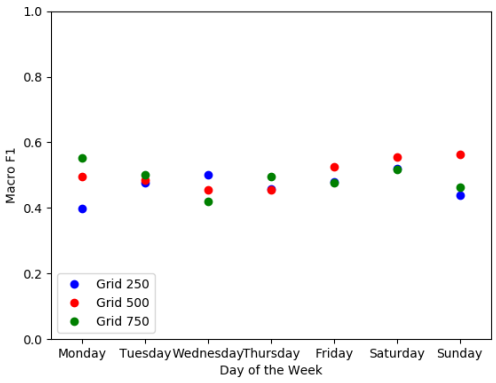
The results for the feature importance are displayed in Figures 5.12 (a), (b) and (c) and it shows the average feature importance across all days for each grid size. Displayed in Figure 5.12 (a) are the results for grid size 250*m*. In this grid size, the most important features are the speed and acceleration statistics (1^{st}, 2^{nd}, 3^{rd} quartile and standard deviation, mainly) with highest value belonging to the longitude center. These statistics come mainly from surrounding cells like 16,27 and 19,25. Figure 5.12 (b) shows the results for grid size 500*m* where the two main features are the 1^{st} and 3^{rd} quartiles for cells 9,12 and 8,13 respectively, from intervals 15*min* and 45*min*. Also important is the *hour* feature, and surrounding cells like 8,11 or 6,10 that hold main roads within them. Lastly, Figure 5.12 (c) represents grid size 750*m* where longitude holds the largest



(a) Micro F1



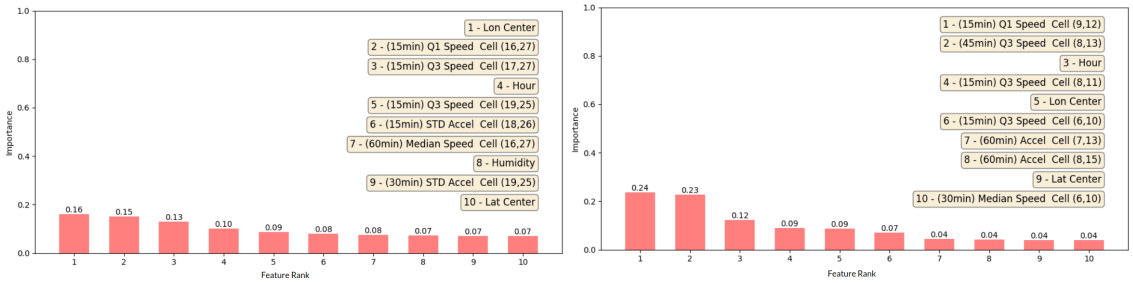
(b) Average F1



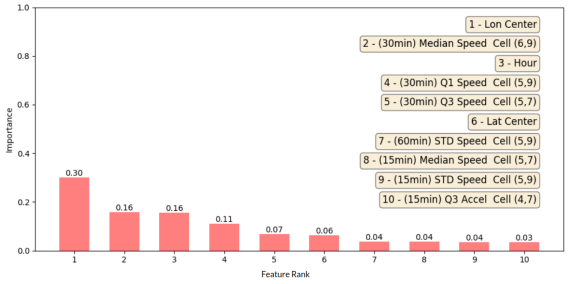
(c) Macro F1

Figure 5.11: Weekday model comparison between all grids.

weight (0.30). Other relevant features are the speed and acceleration statistics from the urban and surrounding cells like 5,9 and 5,7, respectively.



(a) Top 10 Features' Importance, Grid Size 250m - Monday(b) Top 10 Features' Importance, Grid Size 500m - Thursday



(c) Top 10 Features' Importance, Grid Size 750m - Wednesday

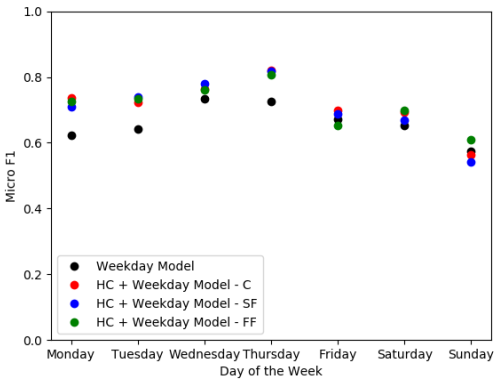
Figure 5.12: Top Features' Importance for all Grid Sizes - Weekday Model.

Lastly, comparing between week days and weekend proves no discernible difference in terms of feature importance. This might be caused because even though weekends have a different traffic pattern when compared to weekdays (see Section 3.3.1), they still have a characteristic pattern, which ends up focusing on the same weekday's features.

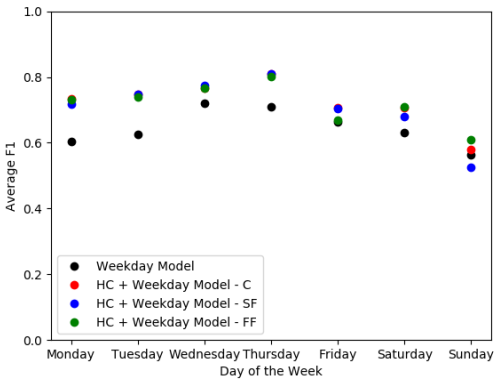
5.2.5 Combination

The last developed model combined the Weekday model with the Hierarchic structure. This allows for a further improvement over the previous model, as two models are created for each day, separating only two classes instead of three, respectively. Figures 5.13 show the results for all metrics for grid size 250m, as an example, with all the possible combinations of macro classes, when compared to the Weekday model.

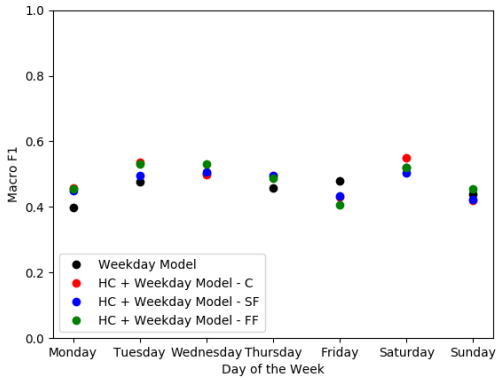
The first thing to notice is how all the metrics had a significant increase for almost all days. Micro F1 in Figure 5.13 (a) increased in Monday, Tuesday and Thursday. It stayed roughly the same in Wednesday, Saturday and Sunday while Friday is the only day where this metric decreased. The increase is justified by an overall increase in the correctly classified time slots across all classes, while the decrease can be explained. For the Macro F1 metric, Figure 5.13 shows it stayed roughly the same throughout the week with just small variations, and it didn't increase its



(a) Micro F1



(b) Average F1

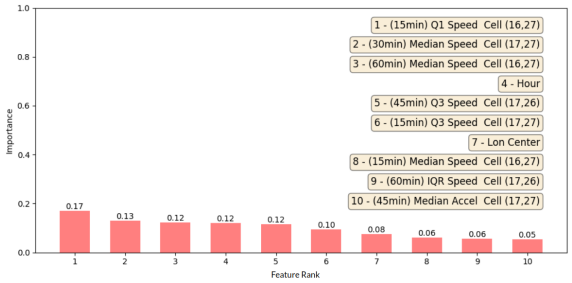


(c) Macro F1

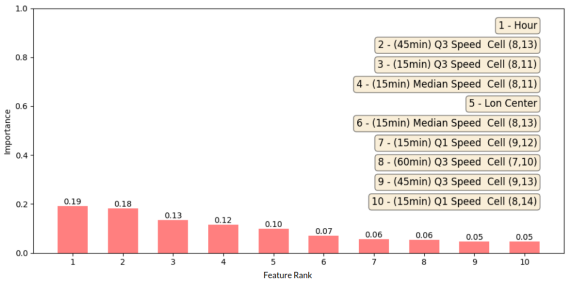
Figure 5.13: Hierarchical and Weekday model Combination comparison with Weekday model for grid size 250m and all macro class combinations. HC refers to he Hierarchical Model, C stands for Congestion, SF is Sync Flow and FF is the Free Flow state.

values in a significant way when compared to Weekday model in the previous section (see Figure 5.11 (c)), which can be caused by at least one class being heavily misclassified.

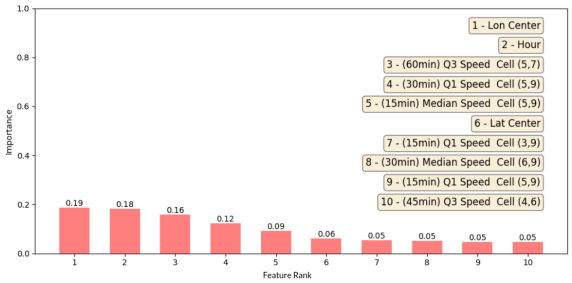
Focusing now on the features for each model and their importance, the results are displayed on Figures 5.14. Across all grid sizes the overall idea seems to be that some specific cells surrounding the urban center have features with higher importance. For grid size 250m this happens for cells 16,27, 17,27 and 17,26, where these features occupy almost the full rank 10. In grid size 500m the main cells are 8,13, 8,11 and 9,12, with the first cell being the exception as it is the main urban cell, the rest of the cells in the Figure are in surrounding areas. Lastly, on grid size 750m, aside from *latitude* and *hour*, the most important features are from 5,9 and 5,7. The first is an urban cell while the second is a surrounding one.



(a) Grid size 250m



(b) Grid size 500m



(c) Grid size 750m

Figure 5.14: Top features' for all grid sizes - Hierarchical and Weekday model combination.

5.2.6 Overall results

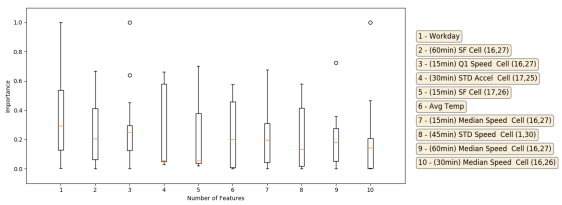
After the analysis of each model, an overall comparison between grid sizes follows in table 5.7 where it is shown the median results of Micro F1, Average F1 and Macro F1 for every different grid size. The table shows that there are no discernible differences between different sizes, as the results are very close across all metrics.

Looking at feature importance, the distribution for the top 10 features of each grid size are displayed in Figures 5.15. The features selected have the highest average importance value across each grid size. The overall idea is that the features which compose the bulk of the most important are the speed and acceleration statistics of the cells that are not in the urban center (e.g. speed's 1^{st} quartile of cell 16,27 for grid size 250m, speed's 1^{st} quartile of cell 8,15 for grid size 500m and

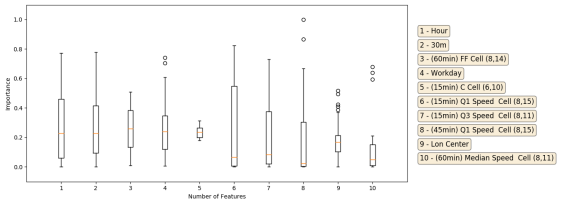
Metric	Grid Size 250m	Grid Size 500m	Grid Size 750m
Micro F1	0.61	0.63	0.62
Average F1	0.61	0.63	0.62
Macro F1	0.41	0.45	0.44

Table 5.7: Overall metrics comparison for all grid sizes.

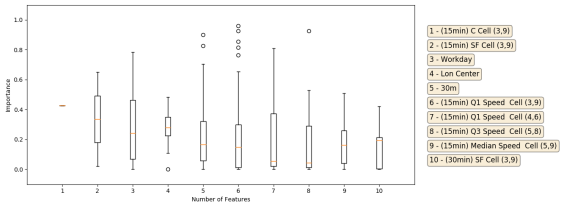
speed's 1^{st} quartile of cell 3,9 for grid size 750m). But even though these features are the most common, they are not the most important. This seems to be the traffic states of the surrounding cells (e.g. *synchronized flow* of cell 16,27 for grid size 250m, *free flow* state for cell 8,14 of grid size 500m and *congestion* and *synchronized flow* of cell 3,9 for grid size 750m) as well as some specific features like the *hour* and *minutes* of the day (grid size 500m) or if the day predicted is a weekend or workday (grid size 250m).



(a) Grid size 250m



(b) Grid size 500m



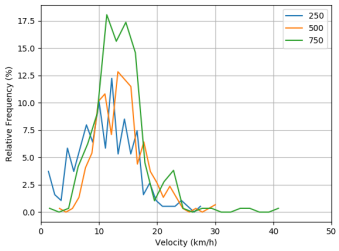
(c) Grid size 750m

Figure 5.15: Top 10 features' box plots for all grid sizes.

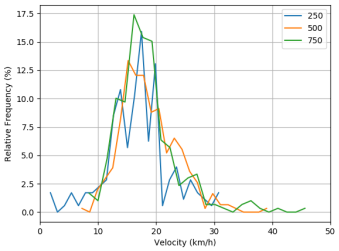
Presented in Figures 5.16 are the values that some of the features take when they appear in the decision trees. Figures 5.16 (a), (b) and (c) represent the variation of the 1^{st}, 2^{nd} and $3r^{rd}$ quartiles for the previous time slot of 15 minutes. For all the grid sizes, the thresholds tend to increase its frequency on roughly the same values: around 13 km/h for the 1^{st} quartile, close to 17 in the 2^{nd} and from 20 to 25 km/h for the 3^{rd} quartile. This shows that the same values for these features are important throughout all the grid sizes available.

In Figure 5.16 (d), the distribution of the *Hour* feature is shown, and again, for the all the grid sizes, the values for this feature is grouped on the same range of hours: in the early morning (from 6 to 8h) and in the afternoon (17,5 to 19,5h). This corresponds to reality as these correspond to the main congestion hours of the city.

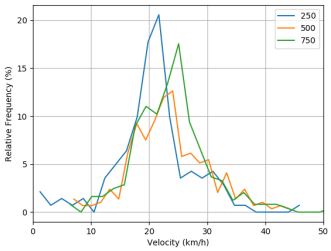
Looking at figure 5.17, the importance of each cell is displayed, when considering only cell



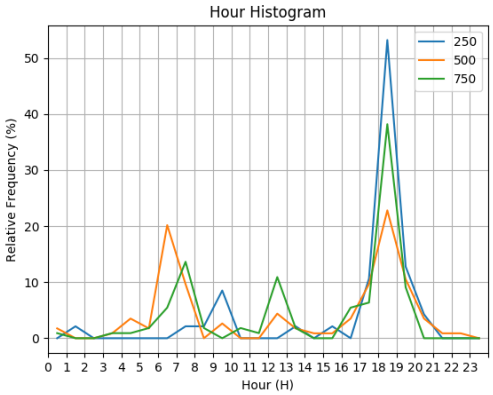
(a) Distribution of the 1^{st} speed quartile, for the first previous time slot



(b) Distribution of the 2^{nd} speed quartile, for the first previous time slot



(c) Distribution of the 3^{rd} speed quartile, for the first previous time slot



(d) Distribution for the Hour feature.

Figure 5.16: Feature Thresholds Variation for all grid sizes.

scope models. These results were obtained by summing every feature that belong to model being created, as well as summing for every other same cell on the same model. *Own cells* box plots refer to the importance that each cell has to its own model, while *Other Cells* is the importance that other cells have. *Other Features* is the importance that features such as weather factors, weekdays, etc., have. The Figure shows that typically, other cells have more importance than the itself.

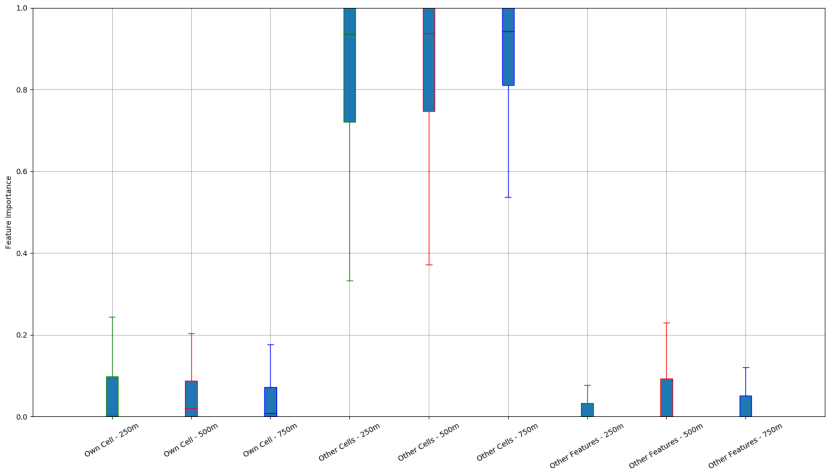


Figure 5.17: Influence of own and neighbor cells.

In Figure 5.18, the importance for each feature from a specific time slot is displayed. Features such as speed or acceleration statistics that belonging to the previous time slot appear on the box plots *15min*. Features that are from 30 minutes before to the slot being predicted are under the *30min* box plots. The rest of the box plots contain features that are from 45 and 60 minutes before the time slot being evaluated. The Figure shows that typically, the features with smaller time slots are more important, most likely because they are the most recent features, and are potentially more similar to the current time slot being predicted. With the small exception of grid size 250m this interpretation can be retrieved from grid sizes 500 and 750m.

Lastly in figure 5.19, the critical cells are shown. These cells were selected by using all the features of every model and calculating the sum of the importance for each cell, across all grid sizes. In the end, only those cells with a total feature importance greater than 0.15 were selected. This resulted in the following cells:

- Grid Size 250m: cell 16,27;
- Grid Size 500m: cells 6,10 and 8,15;
- Grid Size 750m: cells 3,9 and 5,9;

These cells are not located in the urban center nor are the ones with the highest amount of information. They are instead the cells with some of the main roads that lead to the urban center within them, proving these roads are quite important when predicting the traffic state for that area.

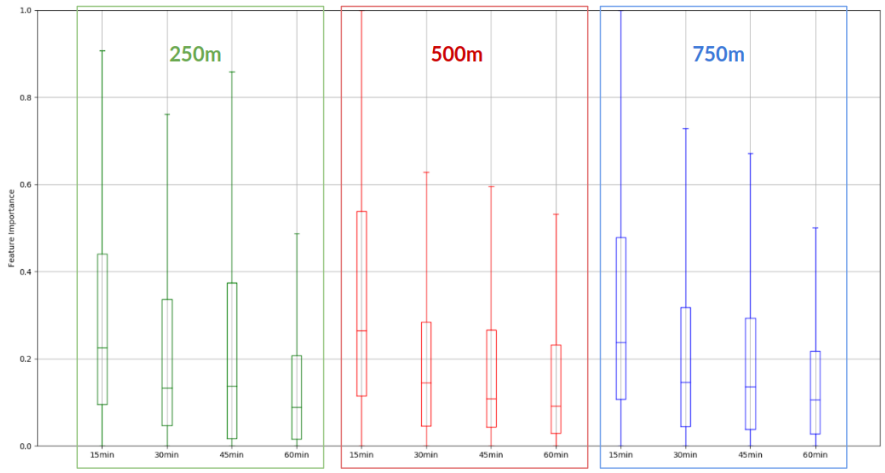


Figure 5.18: Feature importance of the previous time slots for all grid sizes.

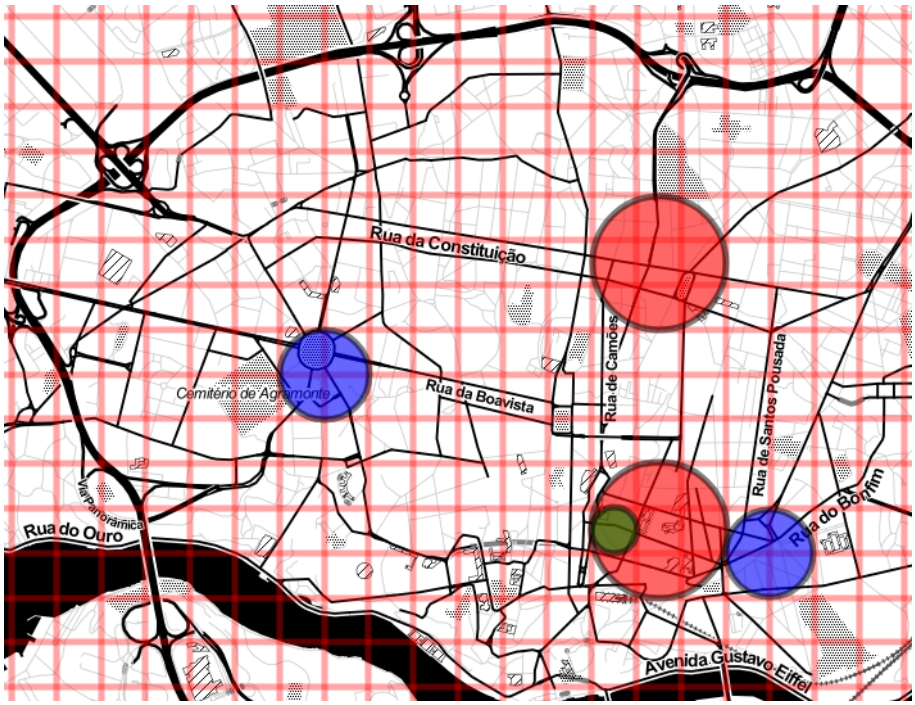


Figure 5.19: Critical Cells.

5.2.7 Method comparison

Since the focus of this project is interpretability rather than prediction, the focus was never to reach the highest possible value for the results of the F1 scores but instead to improve the information retrieved. Even though, when comparing the approach of using a white-box classifier like *Decision Trees* to a method more suited to prediction, like *Random Forests* the results have almost no difference. The results in table 5.8 show that *decision trees* does not stay far behind the results of a *Random Forest* approach, even surpassing it for some cases.

	Grid Size 250*m*		Grid Size 500*m*		Grid Size 750*m*	
Metrics	Decision Tree	Random Forest	Decision Tree	Random Forest	Decision Tree	Random Forest
Micro F1	0.71	0.74	0.72	0.72	0.70	0.70
Average F1	0.72	0.70	0.72	0.71	0.72	0.66
Macro F1	0.49	0.42	0.52	0.46	0.51	0.44
Running Time (s)	1459	3104	2135	6264	2118	5959

Table 5.8: Method comparison - Decision Trees vs. Random Forest.

Table 5.8 displays the median results for each grid and metric and it shows that this project's approach does not fall behind *random forests*, having a score of *Macro F1* always higher and staying on the same level for the other metrics.

Chapter 6

Conclusions

A new approach in terms of congestion estimation is proposed in this thesis. An approach based on interpretability instead of predictability. In order to achieve this, different steps were implemented. First, the removal of erroneous data with a Hampel Filter as well as the implementation of length and duration filters to each of the trips. Next, a series of discretizations were made: space, time and state. Space discretization allowed the city to be studied in small discretized zones, focusing on the top 10 cells with the most information. Three different values are used - 250, 500 and 750m. Next, on time discretization, the data was aggregated in time slots of 15 *min*, ensuring a sufficient amount of information per cell, per time slot. On state discretization, each velocity is classified as one of three traffic states: *congestion*, *synchronized flow* and *congestion*. These would then be used as the ground truth when evaluating each of the developed models.

The next step was the creation of a dataset with all the features available as well as new, calculated, such as 1^{st}, 2^{nd} and 3^{rd} quartiles of both speed and acceleration. Also, the last four states of the surrounding cells were added, which would help to discover if previous traffic states had any influence in the present traffic state.

The following step was the development of different models, that aggregated and separated the data in different ways - Conventional, Weekday, Hierarchical, Sliding Window and a combination of Hierarchical and Weekday - as well in different scopes - global and per cell. Lastly, each model is evaluated using the chosen metrics - Micro F1, Average F1, Macro F1 and Feature Importance. Following an analysis of the generated decision trees, it is shown that the main features that influence the traffic states of the urban area are the speed and acceleration of the surrounding cells. It is also found that grid sizes holds no major influence over the selected features, as no grid size provides more information than the others.

6.1 Future work

For the future work, some different techniques and approaches could be implemented as well as a better management of data. Examples follow, but are not limited to:

- Fundamental Diagram Discretization - Instead of using fixed velocity limits for all cells and across all grid sizes, a fundamental diagram could be inferred for each grid size, or even for each cell, allowing the deduction of important parameters like critical velocities, that would help make a better separation between traffic states;
- Better choosing of cells to create the predicting models - As shown by this project, there are cells that are important by their location and not only by their amount of information. This could be used as a new selection criteria to which cells to build the models on, or to retrieve more information from them.
- Increase the amount of data - For various reasons, the amount of used data was of one month. If this period were to be increased to six months or even a full year, this could give chance for certain features like different weather statistics, to gain more weight. This would happen as there would be more variation from these features throughout the year with the different seasons, maybe creating a relationship with the different traffic states, which would increase the features' weight.
- Increased Point of Interest information - As similar to other projects (Zhan et al., 2017), information about certain points of interest could increase the results, as well as give another insight to traffic's congestion. Information such as the location of roundabout's, schools, or shopping malls, has proven to increase the models and could help to further understand how traffic flows through a city.

References

HCM2000: Highway Capacity Manual. Transportation Research Board, 2000.

Ali Abdelfattah and Ata Khan. Models for predicting bus delays. *Transportation Research Record: Journal of the Transportation Research Board*, (1623):8–15, 1998.

Lukas Ambühl and Monica Menendez. Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Research Part C: Emerging Technologies*, 71:184–197, 2016.

Lukas Ambühl, Allister Loder, Michiel CJ Bliemer, Monica Menendez, and Kay W Axhausen. Introducing a re-sampling methodology for the estimation of empirical macroscopic fundamental diagrams. *Arbeitsberichte Verkehrs-und Raumplanung*, 1271, 2017.

ASTM. Standard specification for highway weigh-in-motion (wim) systems with user requirements and test methods. E1318-02, 2002.

G. Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017. doi: 10.1016/j.compenvurbsys.2017.05.004.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984. ISBN 9780412048418. URL https://books.google.pt/books?id=JwQx-WOmSyQC.

Leo Breiman. Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, 1996.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. From taxi gps traces to social and community dynamics: A survey. *ACM Comput. Surv.*, 46(2):17:1–17:34, December 2013a. ISSN 0360-0300. doi: 10.1145/2543581.2543584. URL http://doi.acm.org/10.1145/2543581.2543584.

Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. From taxi gps traces to social and community dynamics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):17, 2013b.

C.Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 197–203, Oct 2008. doi: 10.1109/ITSC.2008.4732534.

Sing Yiu Cheung and Pravin Pratap Varaiya. *Traffic surveillance by wireless sensor networks*. PhD thesis, University of California, Berkeley, 2006.

Mario Cools, Elke Moons, and Geert Wets. Assessing the impact of weather on traffic intensity. *Weather, Climate, and Society*, 2(1):60–68, 2010. doi: 10.1175/2009WCAS1014.1. URL https://doi.org/10.1175/2009WCAS1014.1.

Ariane Debyser. Urban mobility - shifting towards sustainable transport systems. *EPRS - European Parliamentary Research Service*, 2014. URL http://www.europarl.europa.eu/RegData/etudes/IDAN/2014/538224/EPRS_IDA(2014)538224_REV1_EN.pdf.

Thierry Derrmann, Raphaël Frank, and Francesco Viti. Towards estimating urban macroscopic fundamental diagrams from mobile phone signaling data: A simulation study. Technical report, 2017.

M. Ferreira and P. M. d'Orey. On the impact of virtual traffic lights on carbon emissions mitigation. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):284–295, March 2012. ISSN 1524-9050. doi: 10.1109/TITS.2011.2169791.

BD Greenshields, Ws Channing, Hh Miller, et al. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board, 1935.

Shin-Ting Jeng and Lianyu Chu. A high-definition traffic performance monitoring system with the inductive loop detector signature technology. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 1820–1825. IEEE, 2014.

Lawrence A Klein, Milton K Mills, and David RP Gibson. Traffic detector handbook: -volume i. Technical report, 2006.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

Xiangjie Kong, Zhenzhen Xu, Guojiang Shen, Jinzhong Wang, Qiuyuan Yang, and Benshi Zhang. Urban traffic congestion estimation and prediction based on floating car trajectory data. *Future Generation Computer Systems*, 61:97 – 107, 2016. ISSN 0167-739X. doi: http://dx.doi.org/10.1016/j.future.2015.11.013. URL http://www.sciencedirect.com/science/article/pii/S0167739X15003611.

Jaimyoung Kwon and Kevin Murphy. Modeling freeway traffic with coupled hmms. Technical report, 2000.

Jaimyoung Kwon, Pravin Varaiya, and Alexander Skabardonis. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. *Transportation Research Record: Journal of the Transportation Research Board*, (1856):106–117, 2003.

Guillaume Leduc. Road traffic data: Collection methods and applications. 01 2008.

Xiying Li, Yongye She, Donghua Luo, and Zhi Yu. A traffic state detection tool for freeway video surveillance system. *Procedia-Social and Behavioral Sciences*, 96:2453–2461, 2013.

Vivek Mehta and Inderveer Chana. *Urban Traffic State Estimation Techniques Using Probe Vehicles: A Review*, pages 273–281. Springer Singapore, Singapore, 2017. ISBN 978-981-10-3935-5. doi: 10.1007/978-981-10-3935-5_28. URL https://doi.org/10.1007/978-981-10-3935-5_28.

Luz Elena Y Mimbela and Lawrence A Klein. Summary of vehicle detection and surveillance technologies used in intelligent transportation systems. 2000.

Erik Minge, Jerry Kotzenmacher, and Scott Peterson. Evaluation of non-intrusive technologies for traffic detection. Technical report, Minnesota Department of Transportation, Research Services Section, 2010.

John Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3(4):319–342, Mar 1989. ISSN 1573-0565. doi: 10.1007/BF00116837. URL https://doi.org/10.1007/BF00116837.

Pitu Mirchandani and Larry Head. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, 9(6):415 – 432, 2001. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/S0968-090X(00)00047-4. URL http://www.sciencedirect.com/science/article/pii/S0968090X00000474.

Fred Moses. Weigh-in-motion system using instrumented bridges. *Journal of Transportation Engineering*, 105(3), 1979.

Ronald K. Pearson, Yrjö Neuvo, Jaakko Astola, and Moncef Gabbouj. Generalized hampel filters. *EURASIP Journal on Advances in Signal Processing*, 2016(1):87, Aug 2016. ISSN 1687-6180. doi: 10.1186/s13634-016-0383-6. URL https://doi.org/10.1186/s13634-016-0383-6.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL https://doi.org/10.1007/978-0-387-39940-9_565.

Meead Saberi, Hani Mahmassani, Tian Hou, and Ali Zockaie. Estimating network fundamental diagram using three-dimensional vehicle trajectories: Extending edie's definitions of traffic flow variables to networks. *Transportation Research Record: Journal of the Transportation Research Board*, (2422):12–20, 2014.

Ralf-Peter Schäfer, Kai-Uwe Thiessenhusen, Elmar Brockfeld, and Peter Wagner. A traffic information system by means of real-time floating-car data. 2002.

Toru Seo, Alexandre M. Bayen, Takahiko Kusakabe, and Yasuo Asakura. Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control*, 43:128 – 151, 2017. ISSN 1367-5788. doi: http://dx.doi.org/10.1016/j.arcontrol.2017.03.005. URL http://www.sciencedirect.com/science/article/pii/S1367578817300226.

Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

Aleksandar Stevanovic, Jelka Stevanovic, Kai Zhang, and Stuart Batterman. Optimizing traffic control to reduce fuel consumption and vehicular emissions: Integrated approach with vissim, cmem, and visgaost. *Transportation Research Record: Journal of the transportation research board*, (2128):105–113, 2009.

Senzhang Wang, Lifang He, Leon Stenneth, Philip S Yu, and Zhoujun Li. Citywide traffic congestion estimation with social media. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 34. ACM, 2015.

Y. C. F. Wang and D. Casasent. A support vector hierarchical method for multi-class classification and rejection. In *2009 International Joint Conference on Neural Networks*, pages 3281–3288, June 2009. doi: 10.1109/IJCNN.2009.5178670.

Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

Xianyuan Zhan, Yu Zheng, Xiuwen Yi, and Satish V Ukkusuri. Citywide traffic volume estimation using trajectory data. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):272–285, 2017.

Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

Tongyu Zhu, Fajin Ma, Tao Ma, and Congcong Li. The prediction of bus arrival time using global positioning system data and dynamic traffic information. In *Wireless and Mobile Networking Conference (WMNC), 2011 4th Joint IFIP*, pages 1–5. IEEE, 2011.

Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang. A compressive sensing approach to urban traffic estimation with probe vehicles. *IEEE Transactions on Mobile Computing*, 12(11):2289–2302, Nov 2013. ISSN 1536-1233. doi: 10.1109/TMC.2012.205.